ACS chemical biology

Articles

pubs.acs.org/acschemicalbiology

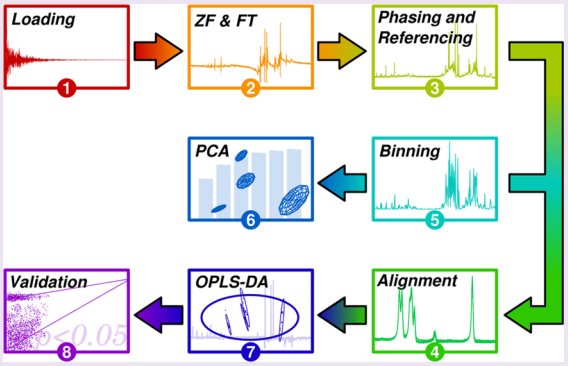
MVAPACK: A Complete Data Handling Package for NMR Metabolomics

Bradley Worley and Robert Powers*

Department of Chemistry, University of Nebraska-Lincoln, Lincoln, Nebraska 68588-0304, United States

Supporting Information

ABSTRACT: Data handling in the field of NMR metabolomics has historically been reliant on either in-house mathematical routines or long chains of expensive commercial software. Thus, while the relatively simple biochemical protocols of metabolomics maintain a low barrier to entry, new practitioners of metabolomics experiments are forced to either purchase expensive software packages or craft their own data handling solutions from scratch. This inevitably complicates the standardization and communication of data handling protocols in the field. We report a newly developed open-source platform for complete NMR metabolomics data handling, MVAPACK, and describe its application on an example metabolic fingerprinting data set.



The biochemical procedures involved in metabolomics experiments are potentially straightforward and inexpensive, depending on the biological systems and pathways under study.[1] The minimal sample handling requirements of one-dimensional (1D) ^{1}H NMR spectroscopy and the immense sensitivity of multivariate statistical methods such as Principal Component Analysis (PCA) and Partial Least Squares (PLS) make NMR metabolic fingerprinting especially attainable. This low barrier to entry has no doubt contributed to the rapid growth of the field. Unfortunately, commercial software packages available for multivariate analysis (SIMCA, PLS Toolbox, The Unscrambler, etc.) tend to be expensive and require more software for upstream processing and treatment of spectral data. Furthermore, such packages provide little to no domain-specific functionality, requiring a user to first open and preprocess NMR data in ACD/1D NMR Manager (Advanced Chemistry Development) or Mnova NMR (Mestrelabs Research) and perform further statistical pretreatment in MATLAB (The MathWorks, Natick, MA) or Microsoft Excel. This results in an unnecessarily cumbersome and time-consuming data handling pipeline by forcing the user to pass data between multiple software packages. As a result, the field of metabolomics research is littered with unpublished "in-house" software solutions created for processing or modeling NMR data sets.[2−8] This continued reinvention of the wheel impedes progress in the field and complicates the tasks of standardization and communication of protocols that the metabolomics community is attempting to achieve.[9,10] Unfortunately, these in-house solutions are far less likely than their commercial counterparts to include proper means of validating supervised multivariate models, further contributing to the general lack of model validation currently present in the field.[11] While the community has released several official software packages for metabolomics,[12−18] none provide a complete, well-validated data path. To our knowledge, no single software package exists to bring raw NMR data along its complete journey to validated, interpretable multivariate models.

We have developed a free and open-source software package, MVAPACK, that provides a complete pipeline of functions for NMR chemometrics and metabolomics. MVAPACK is written in the GNU Octave mathematical programming language,[19] which is also open-source and nearly syntactically identical to MATLAB. Thus, the installation of GNU/Linux, Octave, and MVAPACK onto a commodity workstation provides a uniform environment in which a data analyst may truly work "from FIDs to models" in a few minutes using a set of well-documented, open-source, high-level processing functions.

The functions available in MVAPACK span the following general categories: data loading, preprocessing, pretreatment, modeling, and validation.[9] Loading of Bruker data is available using either a high-performance DMX-format loading routine or NMRPipe[20] as a backend, and loading of Agilent data is available using an NMRPipe backend. Additionally, data in a variety of text formats may be read into MVAPACK using standard GNU Octave routines. The preprocessing functions in MVAPACK follow the traditional paradigm of NMR processing and include methods for apodization, zero-filling, Fourier transformation, manual and automatic phase correction,[21,22] region of interest selection, peak picking,[23] integration, and referencing. Functions for data pretreatment in MVAPACK include scaling,[24] normalization,[4,25,26] binning and align-

Received: December 3, 2013
Accepted: February 27, 2014
Published: February 27, 2014

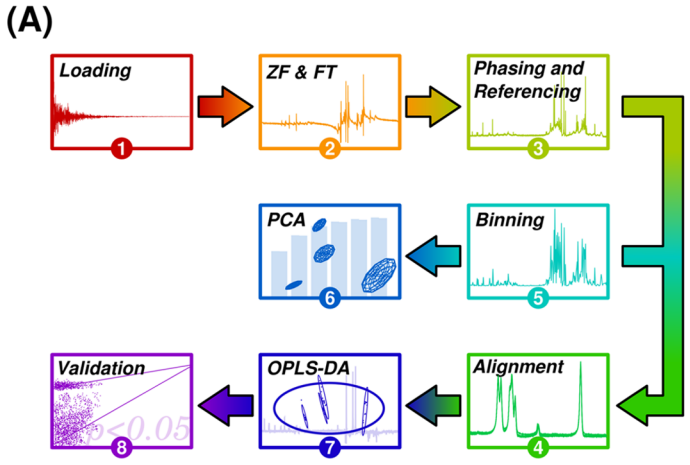


Figure 1. An example NMR metabolic fingerprinting data handling flow diagram (A) and its associated MVAPACK commands (B). This minimalistic data handling script is a simple starting point for using MVAPACK; much more flexibility and functionality are present in the software than can be shown here. All functions in boldface are provided in MVAPACK.

ment,[27−29] and denoising.[30] Finally, MVAPACK provides complete support for building PCA, LDA (Linear Discriminant Analysis), PLS, and OPLS (Orthogonal Projections to Latent Structures) models from processed and treated data sets.[31−35] All models are validated as they are built based on Monte Carlo *n*-fold internal cross-validation,[11,36−39] which is also utilized to determine the number of significant model components. Further validation of supervised models is available in the form of CV-ANOVA[40] and response permutation[11] significance testing.

This work describes the structure of MVAPACK and an application of MVAPACK to a use-case that is representative of many metabolomics studies: the NMR metabolic fingerprinting of coffee for discrimination of four roasts based on either general spectral trends or caffeine concentration.

METHODS

Data Sets. To illustrate the capabilities of MVAPACK on a real experimental data set, four roasts of brewed coffee were purchased from a local coffee shop, and replicate samples were made from each roast. A final set of 64 ^{1}H NMR spectra (N = 64, K = 16384) was obtained and used for PCA, LDA, and OPLS-R multivariate analyses. Estimates of caffeine concentration were also obtained from liquid−liquid extractions of each roast into CH_2Cl_2 followed by UV−vis spectroscopy.[41] See the Supplementary Methods for detailed information about the processing of the Coffees ^{1}H NMR and UV−vis data sets.

Software Implementation. The MVAPACK software package is written in GNU Octave, an open-source mathematical programming language that uses MATLAB syntax. Every function available in MVAPACK is realized as a single Octave function file that may be examined or changed using any text editor. Most functions in MVAPACK follow a similar input-to-output template, where an input data matrix ***A*** is modified and returned as an output data matrix ***B***. Other required input arguments may accompany ***A***, and extra output values may accompany ***B***, depending on the requirements of the user. Furthermore, models produced by PCA, PLS, OPLS, and LDA are all similarly organized into Octave structures that all follow scalar, vector, and matrix notations of Wold et al.[35] Thus, functions in MVAPACK are highly modular, often allowing drop-in replacement of one processing or modeling algorithm for another by a simple change of function name and arguments.

Data may be handled by MVAPACK in either interactive mode, in which the user types commands into the Octave interpreter one at a time, or as a script, where a complete processing scheme has been laid out in an Octave script to be executed noninteractively. Once an ideal set of processing commands and parameters is determined by interactive manipulation of the data, it may be immortalized in an Octave script, thus providing documentation of procedures and allowing for rapid recalculation of all associated results.

Figure 1 illustrates a simple MVAPACK script capable of taking 1D ^{1}H NMR data from free induction decays to validated PCA and OPLS-DA models. In section 1, a binary class matrix ***Y*** and an accompanying set of class labels are built, and the time-domain data is loaded into the data matrix ***F***. In section 2, the time-domain data matrix ***F*** is zero-filled once and Fourier transformed to produce the spectral data matrix ***S***. Section 3 automatically phase corrects the spectra in ***S***, normalizes and corrects between-spectrum phase differences, and corrects the chemical shift abscissa to center the reference peak at 0 ppm. In sections 4 and 5, processing splits into two pathways, where icoshift alignment[27] is used to generate a data matrix fit for full-resolution OPLS-DA (***A***) and optimized binning[28] is used to generate a data matrix for PCA (***B***). In section 6, a PCA model is built and assigned classes and labels, and a model quality plot and a scores plot are produced. In section 7, similar functions are used to build an OPLS-DA model and produce summary plots. Finally, section 8 performs CV-ANOVA[40] and response permutation[11] significance tests to fully validate the supervised OPLS-DA model. While Figure 1 is complete, it is still an extremely bare-bones approach to metabolic fingerprinting. MVAPACK provides countless other functions and schemes for processing data. Detailed information about all MVAPACK functionality is available in the MVAPACK manual online.

Software Validation. Validation of the proper operation of the NMR processing functions of MVAPACK was performed by visually comparing the MVAPACK-processed 1D ^{1}H NMR spectra from the Coffees data set (Figure 2) with the processed NMR spectra produced by ACD/1D NMR Manager (Advanced Chemistry Development).

Verification of icoshift alignment performance was performed using the Wine ^{1}H NMR data set[42] available from the University of Copenhagen. As this data set contains large amounts of chemical shift dispersion due to differences in chemical properties of each wine, it is an ideal basis for assessing the performance of NMR peak alignment algorithms (Figure 3).

Validation of the proper operation of PCA, PLS and OPLS multivariate decompositions was performed by comparing the scores produced by analysis of the Coffees NMR data set in MVAPACK with those produced by SIMCA-P+ 13.0 (Umetrics AB, Umea, Sweden) (Figures 4 and 5).

dx.doi.org/10.1021/cb4008937 | *ACS Chem. Biol.* 2014, 9, 1138−1144

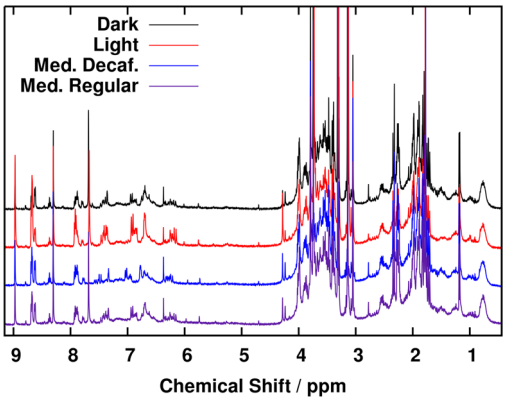


Figure 2. Representative processed 1D ^{1}H NMR spectra for each analyzed coffee roast, acquired using the water-suppressed CPMG-*z* pulse sequence and processed in MVAPACK. To reach this point, free induction decays were simply Fourier transformed and automatically phased. No manual phase corrections were applied after autophasing.

RESULTS AND DISCUSSION

Results. Use of MVAPACK during analysis of the coffees data set arguably facilitated rapid identification of ideal processing and modeling parameters during data handling. Use of automatic phase correction,[21] optimized binning,[28] and PQ normalization[4] yielded a data set in which three principal components were sufficient to fully separate all classes in scores space, and subsequent LDA modeling resulted in complete class separation in only two components (Figure 4). During the process of optimizing the data handling, modifying the procedure required nothing more than changing a few commands in a GNU Octave script, not unlike changing processing parameters in an NMRPipe script, although considerably more human-readable.

As opposed to the PCA modeling, which utilized binned spectra, OPLS-R modeling was performed on full-resolution 1D ^{1}H NMR spectra in order to reap the interpretive advantages of full-resolution backscaled loadings[3] and greater support for each loading 'peak' in S-plots[8] (Figures 5 and 6). The availability of icoshift alignment[27] in MVAPACK effectively makes the modeling of full-resolution NMR spectra possible by correcting positional noise in the spectra that corrupts the bilinear nature of the data (Figure 3). By regressing the NMR data against estimates of caffeine concentration obtained by UV−vis spectroscopy (Supplementary Figure 1S), a loadings pseudospectrum of caffeine was obtained that matched almost perfectly with spectral data deposited in the Biological Magnetic Resonance Bank (Figure 6).[43] It is conceivable that spectral features coextracted with caffeine in the loadings correspond to coffee bean metabolites lost alongside caffeine during roasting or decaffeination.

Notably, the UV−vis-estimated caffeine concentration of the dark roast coffee was slightly higher than that of the medium regular roast, which is contrary to expectation given that the coffees were brewed using equal volumes of grounds. However, OPLS-R of the NMR data using the estimated caffeine concentrations correctly ranked the roasts according to expectation. When more orthogonal components were allowed into the OPLS-R model, the dark roast again shifted to a higher caffeine concentration, beautifully indicating the presence of slight overfitting (data not shown). Therefore, an OPLS-R model having only a single orthogonal component was chosen, given the fact that it more faithfully modeled the underlying

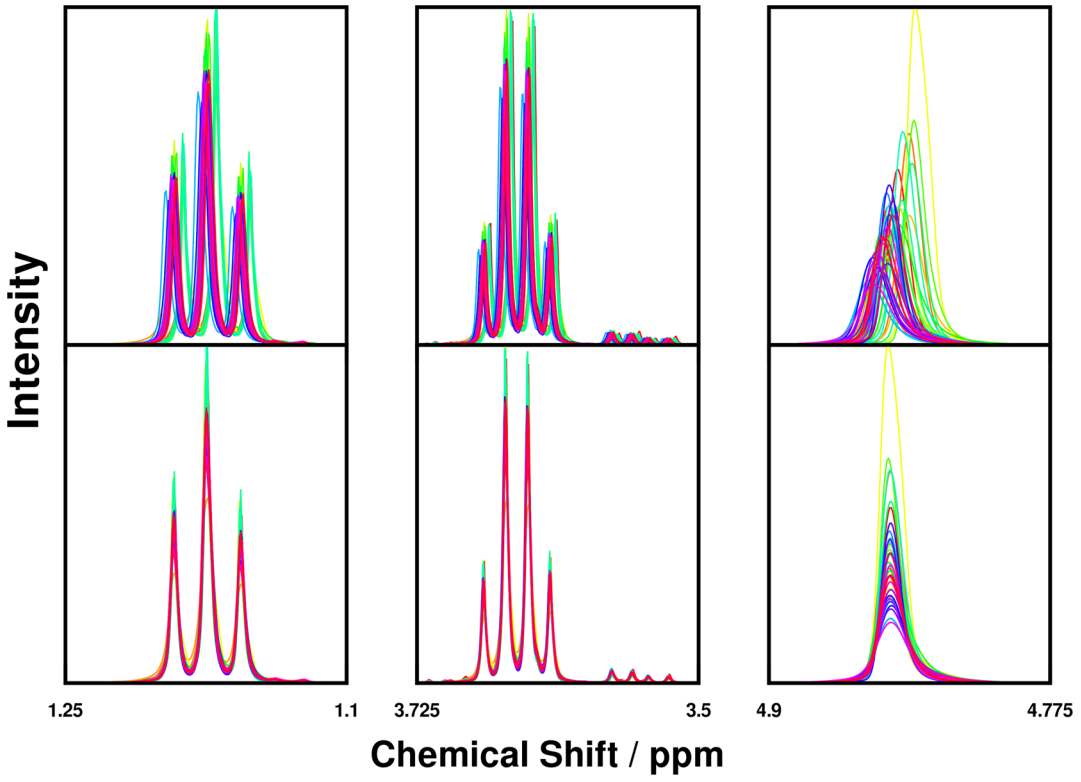


Figure 3. Comparison between the raw (upper) and Interval Correlation Optimized Shifted (*icoshift*, lower) alignment of the wines data set, showing the resulting alignment of the three major spectral features (ethanol H_{C2}, left; ethanol H_{C1}, middle; residual water, right).

dx.doi.org/10.1021/cb4008937 | *ACS Chem. Biol.* 2014, 9, 1138−1144

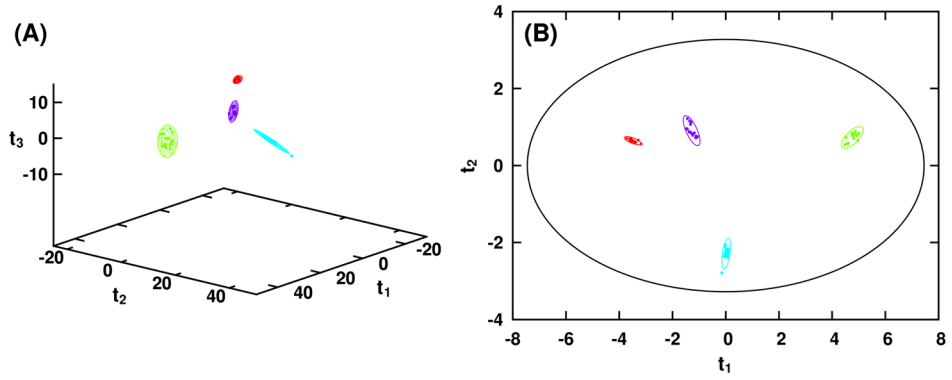


Figure 4. PCA (A) and LDA (B) scores of the four coffee roasts. Red points represent dark roast, green points represent light roast, cyan points represent medium decaffeinated roast, and violet points represent medium regular roast. Ellipsoids and ellipses enclose the 95% confidence intervals estimated by the sample means and covariances of each class. Note that the axis labels in panels A and B indicate scores in PCA and LDA bases, respectively, and not the same set of scores. The PCA internal cross-validation results are summarized in Supplementary Figure 2S, and the LDA response permutation testing results are summarized in Supplementary Figure 4S.

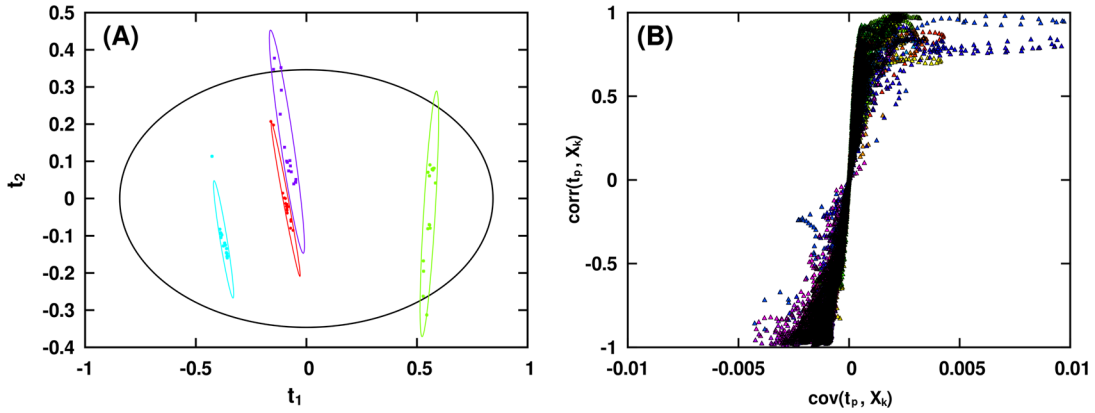


Figure 5. OPLS-R scores plot (A) and S-plot (B) of the four coffee roasts, where each coffee roast was regressed against its caffeine concentration estimated by UV−vis. Points and ellipses in the scores plot follow the same color scheme to those in Figure 4. Spectral variables in the upper right quadrant of the S-plot correspond to caffeine NMR resonances. The internal cross-validation results are summarized in Supplementary Figure 3S, and the response permutation testing results are summarized in Supplementary Figure 5S.

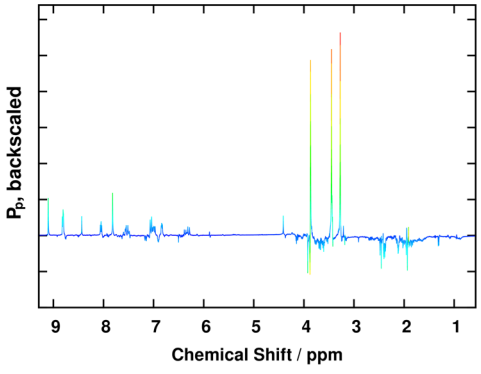


Figure 6. Backscaled OPLS-R predictive loadings of the four coffee roasts regressed according to estimated caffeine concentration. The pseudospectral nature of backscaled loadings facilitates analysis of model results by any spectroscopist. The four most intense positive peaks in the loadings pseudospectrum correspond directly to caffeine NMR resonances archived in the BMRB, indicating a fairly successful regression against caffeine concentration.

NMR data at the expense of contradicting the more uncertain UV−vis measurements.

Finally, no discernible difference was observed between the 1D ^{1}H NMR spectra acquired with and without T_2-filtering. Spectra collected on in-house brewed coffee exhibited high levels of protein background signal, which were readily suppressed using the CPMG-z pulse sequence element. On the other hand, the spectra of the four purchased roasts showed no such background signal, possibly due to more correct brewing technique.

Discussion. We have presented MVAPACK, a completely free and open-source data handling environment for NMR chemometrics targeted toward 1D ^{1}H NMR metabolic fingerprinting applications, and described its use on a representative data set of four coffee roasts to identify discriminating spectral features and chemical trends. Unlike data handling tool chains composed of multiple commercial software packages, MVAPACK is free to use, modify, and distribute according to the GNU General Public License and provides a single consistent data handling environment. Because MVAPACK is written for GNU Octave, researchers already familiar with MATLAB syntax will also be familiar with MVAPACK without a considerable learning curve. Data sets and results obtained using MVAPACK are readily saved and exchanged using GNU Octave built-in support for the MATLAB MAT-file format.

A recent review[44] of software packages targeted at metabolomics highlights the piecemeal nature of 1D ^{1}H NMR data handling in the field, where no single software

dx.doi.org/10.1021/cb4008937 | ACS Chem. Biol. 2014, 9, 1138−1144

package is capable of performing all the tasks required by the analyst (Supplementary Table S1). MVAPACK addresses this need by providing a complete pipeline that is tuned for metabolic fingerprinting. Use of MVAPACK reduces data analysis time in metabolic fingerprinting from days to minutes, simply by collecting all the required processing and modeling functions into a single scriptable environment. In fact, the example script in Figure 1 would execute in under 5 min on a modern GNU/Linux or Mac OS X computer system.

The routine processing of *any* 1D NMR spectral data may be readily done with MVAPACK. As illustrated in Figure 2, the processing of the Coffees NMR data set with MVAPACK yielded an outcome consistent with any commercial or standardized NMR processing suite. Moreover, processing routines are easily batched. The MVAPACK script written to automate the rapid processing and modeling of the Coffees NMR data set was composed of intuitive, modular commands that logically subdivide the script into recognizable tasks like automatic phase correction, referencing, etc. Furthermore, aside from physical memory limitations of the host computer, MVAPACK does not impose any limit in the number of NMR FIDs that may be simultaneously processed.

NMR spectral data presents a unique challenge to multivariate statistical algorithms due to chemical shift variations between spectra caused by differences in temperature, pH, ionic strength, chemical exchange, etc. These variations "blur" true spectral correlations across multiple variables, resulting in lower quality models from linear methods like PCA and PLS.[3,45] To address this problem, chemometric treatments of NMR data include either a binning or alignment procedure to numerically mask or synthetically correct, respectively, peak misalignment. MVAPACK provides tested implementations of both an optimized binning algorithm (OBA) as described by Sousa et al.,[28] an adaptive binning method described by De Meyer et al.,[29] and the icoshift alignment algorithm.[27] The OBA and AI-binning methods minimize the splitting of peaks between multiple bins and significantly reduce the size of the data matrix, thus reducing PCA computational time. Conversely, icoshift maintains the original dimensionality of the data set and allows for the possibility of generating backscaled OPLS loadings that greatly enhance overall model interpretability (Figure 6). The implementation of the icoshift algorithm within MVAPACK was evaluated against the Wine 1D ^{1}H NMR data set,[42] which exhibits substantial peak position variability due to pH and ionic strength differences between each wine. Figure 4 shows the results of MVAPACK icoshift alignment of the major spectral features present in the Wine data set. It is evident that the MVAPACK implementation of icoshift performs on par with published results from the existing implementation by Savorani et al.[27] Similarly, MVAPACK includes a wide variety of normalization, scaling, and denoising methods routinely used by the metabolomics community for pretreatment of NMR data sets. This includes our recently described phase scatter correction (PSC) normalization method, which has been shown to outperform previous methods in applications requiring PCA or PLS decomposition of NMR spectral data.[46]

While no two metabolomics data sets are created equally, we have identified and highlighted a core set of functions in MVAPACK that serves as an optimal starting point when processing and modeling 1D ^{1}H NMR data sets (Figure 1). Use of minimal time-domain processing functions, automatic phase correction combined with PSC normalization, and basic referencing can often yield a routinely reproducibly processed data set without any analyst intervention. Furthermore, PCA of OBA-binned data combined with OPLS-DA of icoshift-aligned spectra produces an effective balance when both general chemical trends and class-discriminating spectral features are sought. Rigorous validation of supervised models, in the form of CV-ANOVA and permutation testing, adds a necessary level of confidence in the interpretation and reuse of supervised models. In our hands, this core function set provides a sane starting point during the handling of new data sets, from which optimization of processing and treatment is a simple matter of tweaking a script file.

A major advantage of MVAPACK is the seamless transfer of the processed and treated NMR data to multivariate statistical analyses. The PCA, PLS, OPLS, and LDA linear modeling algorithms, now ubiquitous in the metabolomics community, are all implemented in MVAPACK. Model results may be visualized and interpreted using MVAPACK routines that provide scatter and line plots of model scores and loadings in a variety of forms. Critically, MVAPACK automatically ensures that *all* produced models are valid using *n*-fold Monte Carlo internal cross-validation[37,38] routines and provides further means of validating supervised models in the forms of CV-ANOVA[40] and response permutation[11] significance testing (Supplementary Figures 2S−5S). The Coffees NMR data set was used to provide a demonstration of the capabilities of MVAPACK when applied to real metabolomics data. The resulting PCA, LDA and OPLS-R scores and the OPLS-R S-plot are depicted in Figures 4 and 5. SIMCA-P+ was also used to generate the same set of scores from the Coffees NMR data set. A comparison of the PCA and OPLS-R scores between MVAPACK and SIMCA-P+ is shown in Supplementary Figures 6S and 7S. Exact agreement was found between all models' scores to within the numerical precision available from SIMCA-P+. Because it implements well-established algorithms available from peer-reviewed chemometrics literature, MVAPACK generates identical results compared to an expensive commercial software package (SIMCA-P+) that is arguably the standard in multivariate data analysis.

In short, MVAPACK provides a complete platform for NMR chemometrics data handling that is ideal for both routine handling of metabolomics data sets and development of novel chemometrics algorithms. Unlike its closed-source predecessors, the modular, open-source design of MVAPACK readily accepts new functionality, allowing it to grow and maintain pace with the state-of-the-art in the chemometrics field. MVAPACK is freely available for download at http://bionmr.unl.edu/mvapack.php. Detailed documentation of MVAPACK and the presented Coffees data set and all its associated processing scripts and results are also available for download.

■ ASSOCIATED CONTENT

Supporting Information

Supplementary figures and data handling methods related to the coffees data set. Table that compares the NMR and metabolomics software features of MVAPACK with 15 other software packages. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Email: rpowers3@unl.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors would like to thank J. Catazaro, T. Gebregiworgis, and D. Marshall for their assistance in beta-testing and troubleshooting MVAPACK. This project was supported by National Institutes of Health Grants P20 RR-17675, P30 GM103335, R01 CA163649-01A1, and R01 AI087668-01A1. Research was performed in facilities renovated with support from the NIH under Grant RR015468-01.

■ REFERENCES

(1) 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. *J. Integr. OMICS 3*, 120−137.

(2) Cloarec, O., Dumas, M. E., Craig, A., Barton, R. H., Trygg, J., Hudson, J., Blancher, C., Gauguier, D., Lindon, J. C., Holmes, E., and Nicholson, J. (2005) Statistical total correlation spectroscopy: An exploratory approach for latent biomarker identification from metabolic H-1 NMR data sets. *Anal. Chem. 77*, 1282−1289.

(3) Cloarec, O., Dumas, M. E., Trygg, J., Craig, A., Barton, R. H., Lindon, J. C., Nicholson, J. K., and Holmes, E. (2005) Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in H-1 NMR spectroscopic metabonomic studies. *Anal. Chem. 77*, 517−526.

(4) 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.

(5) Kang, J., Choi, M., Kang, S., Kwon, H., Wen, H., Lee, C. H., Park, M., Wiklund, S., Kim, H. J., Kwon, S. W., and Park, S. (2008) Application of a H-1 Nuclear Magnetic Resonance (NMR) metabolomics approach combined with orthogonal projections to latent structure-discriminant analysis as an efficient tool for discriminating between Korean and Chinese herbal medicines. *J. Agric. Food Chem. 56*, 11589−11595.

(6) Verhoeckx, K. C. M., Bijlsma, S., Jespersen, S., Ramaker, R., Verheij, E. R., Witkamp, R. F., van der Greef, J., and Rodenburg, R. J. T. (2004) Characterization of anti-inflammatory compounds using transcriptomics, proteomics, and metabolomics in combination with multivariate data analysis. *Int. Immunopharmacol. 4*, 1499−1514.

(7) Viant, M. R. (2003) Improved methods for the acquisition and interpretation of NMR metabolomic data. *Biochem. Biophys. Res. Commun. 310*, 943−948.

(8) Wiklund, S., Johansson, E., Sjostrom, L., Mellerowicz, E. J., Edlund, U., Shockcor, J. P., Gottfries, J., Moritz, T., and Trygg, J. (2008) Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Anal. Chem. 80*, 115−122.

(9) Goodacre, R., Broadhurst, D., Smilde, A. K., Kristal, B. S., Baker, J. D., Beger, R., Bessant, C., Connor, S., Calmani, G., Craig, A., Ebbels, T., Kell, D. B., Manetti, C., Newton, J., Paternostro, G., Somorjai, R., Sjostrom, M., Trygg, J., and Wulfert, F. (2007) Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics 3*, 231−241.

(10) Lindon, J. C., Nicholson, J. K., Holmes, E., Keun, H. C., Craig, A., Pearce, J. T. M., Bruce, S. J., Hardy, N., Sansone, S. A., Antti, H., Jonsson, P., Daykin, C., Navarange, M., Beger, R. D., Verheij, E. R., Amberg, A., Baunsgaard, D., Cantor, G. H., Lehman-McKeeman, L., Earll, M., Wold, S., Johansson, E., Haselden, J. N., Kramer, K., Thomas, C., Lindberg, J., Schuppe-Koistinen, I., Wilson, I. D., Reily, M. D., Robertson, D. G., Senn, H., Krotzky, A., Kochhar, S., Powell, J., van der Ouderaa, F., Plumb, R., Schaefer, H., and Spraul, M. (2005) Summary recommendations for standardization and reporting of metabolic analyses. *Nat. Biotechnol. 23*, 833−838.

(11) Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., Vis, D. J., Smilde, A. K., van Velzen, E. J. J., van Duijnhoven, J. P. M., and van Dorsten, F. A. (2008) Assessment of PLS-DA cross validation. *Metabolomics 4*, 81−89.

(12) Wang, T., Shao, K., Chu, Q. Y., Ren, Y. F., Mu, Y. M., Qu, L. J., He, J., Jin, C. W., and Xia, B. (2009) Automics: an integrated platform for NMR-based metabonomics spectral processing and data analysis. *BMC Bioinf. 10*, 1 DOI: 10.1186/1471-2105-10-83.

(13) Alonso, A., Rodriguez, M. A., Vinaixa, M., Tortosa, R., Correig, X., Julia, A., and Marsal, S. (2014) Focus: a robust workflow for one-dimensional NMR spectral analysis. *Anal. Chem. 86*, 1160−1169.

(14) Izquierdo-Garcia, J. L., Rodriguez, I., Kyriazis, A., Villa, P., Barreiro, P., Desco, M., and Ruiz-Cabello, J. (2009) A novel R-package graphic user interface for the analysis of metabonomic profiles. *BMC Bioinf. 10*, 1 DOI: 10.1186/1471-2105-10-363.

(15) Jarvis, R. M., Broadhurst, D., Johnson, H., O'Boyle, N. M., and Goodacre, R. (2006) PYCHEM: a multivariate analysis package for python. *Bioinformatics 22*, 2565−2566.

(16) Gaude, E., Chignola, F., Spiliotopoulos, D., Spitaleri, A., Ghitti, M., Garcia-Manteiga, J. M., Mari, S., and Musco, G. (2013) muma, an R package for metabolomics univariate and multivariate statistical analysis. *Curr. Metabolomics 1*, 180−189.

(17) Xia, J. G., Mandal, R., Sinelnikov, I. V., Broadhurst, D., and Wishart, D. S. (2012) MetaboAnalyst 2.0-a comprehensive server for metabolomic data analysis. *Nucleic Acids Res. 40*, W127−W133.

(18) Daszykowski, M., Serneels, S., Kaczmarek, K., Van Espen, P., Croux, C., and Walczak, B. (2007) TOMCAT: A MATLAB toolbox for multivariate calibration techniques. *Chemom. Intell. Lab. Syst. 85*, 269−277.

(19) Eaton, J. W., Bateman, D., and Hauberg, S. (2008) *GNU Octave Manual Version 3*, Network Theory Limited, U.K..

(20) Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J., and Bax, A. (1995) NMRPipe - a multidimensional spectral processing system based on unix pipes. *J. Biomol. NMR 6*, 277−293.

(21) Chen, L., Weng, Z. Q., Goh, L. Y., and Garland, M. (2002) An efficient algorithm for automatic phase correction of NMR spectra based on entropy minimization. *J. Magn. Reson. 158*, 164−168.

(22) Siegel, M. M. (1981) The use of the modified simplex-method for automatic phase correction in Fourier-Transform Nuclear Magnetic-Resonance Spectroscopy. *Anal Chim Acta-Comp 5*, 103−108.

(23) Du, P., Kibbe, W. A., and Lin, S. M. (2006) Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics 22*, 2059−2065.

(24) van den Berg, R. A., Hoefsloot, H. C. J., Westerhuis, J. A., Smilde, A. K., and van der Werf, M. J. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics 7*, 142.

(25) Barnes, R. J., Dhanoa, M. S., and Lister, S. J. (1989) Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc. 43*, 772−777.

(26) Torgrip, R. J. O., Aberg, K. M., Alm, E., Schuppe-Koistinen, I., and Lindberg, J. (2008) A note on normalization of biofluid 1D H-1 NMR data. *Metabolomics 4*, 114−121.

(27) 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.

(28) Sousa, S. A. A., Magalhaes, A., and Ferreira, M. M. C. (2013) Optimized bucketing for NMR spectra: Three case studies. *Chemom. Intell. Lab. Syst. 122*, 93−102.

(29) De Meyer, T., Sinnaeve, D., Van Gasse, B., Tsiporkova, E., Rietzschel, E. R., De Buyzere, M. L., Gillebert, T. C., Bekaert, S., Martins, J. C., and Van Criekinge, W. (2008) NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm. *Anal. Chem. 80*, 3783−3790.

(30) Westerhuis, J. A., de Jong, S., and Smilde, A. K. (2001) Direct orthogonal signal correction. *Chemom. Intell. Lab. Syst. 56*, 13−25.

(31) Bylesjo, M., Rantalainen, M., Cloarec, O., Nicholson, J. K., Holmes, E., and Trygg, J. (2006) OPLS discriminant analysis:

combining the strengths of PLS-DA and SIMCA classification. *J. Chemom. 20*, 341−351.
(32) Härdle, W., and Simar, L. (2012) *Applied Multivariate Statistical Analysis*, 3rd ed., Springer, Heidelberg; New York.
(33) Jolliffe, I. T. (2002) *Principal Component Analysis*, 2 ed., Springer, New York.
(34) Trygg, J., and Wold, S. (2002) Orthogonal projections to latent structures (O-PLS). *J Chemom. 16*, 119−128.
(35) Wold, S., Sjostrom, M., and Eriksson, L. (2001) PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst. 58*, 109−130.
(36) Shao, J. (1993) Linear-model selection by cross-validation. *J. Am. Stat. Assoc. 88*, 486−494.
(37) Xu, Q. S., and Liang, Y. Z. (2001) Monte Carlo cross validation. *Chemom. Intell. Lab. Syst. 56*, 1−11.
(38) Xu, Q. S., Liang, Y. Z., and Du, Y. P. (2004) Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *J. Chemom. 18*, 112−120.
(39) Eshghi, P. (2014) Dimensionality choice in principal components analysis via cross-validatory methods. *Chemom. Intell. Lab. 130*, 6−13.
(40) Eriksson, L., Trygg, J., and Wold, S. (2008) CV-ANOVA for significance testing of PLS and OPLS (R) models. *J. Chemom. 22*, 594−600.
(41) 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.
(42) Larsen, F. H., van den Berg, F., and Engelsen, S. B. (2006) An exploratory chemometric study of H-1 NMR spectra of table wines. *J. Chemom. 20*, 198−208.
(43) Ulrich, E. L., Akutsu, H., Doreleijers, J. F., Harano, Y., Ioannidis, Y. E., Lin, J., Livny, M., Mading, S., Maziuk, D., Miller, Z., Nakatani, E., Schulte, C. F., Tolmie, D. E., Kent, W. R., Yao, H., and Markley, J. L. (2008) BioMagResBank. *Nucleic Acids Res. 36*, D402−D408.
(44) Izquierdo-Garcia, J. L., Villa, P., Kyriazis, A., del Puerto-Nevado, L., Perez-Rial, S., Rodriguez, I., Hernandez, N., and Ruiz-Cabello, J. (2011) Descriptive review of current NMR-based metabolomic data analysis packages. *Prog. Nucl. Magn. Reson. Spectrosc. 59*, 263−270.
(45) Stoyanova, R., Nicholls, A. W., Nicholson, J. K., Lindon, J. C., and Brown, T. R. (2004) Automatic alignment of individual peaks in large high-resolution spectral data sets. *J. Magn. Reson. 170*, 329−335.
(46) Worley, B., and Powers, R. (2013) Simultaneous phase and scatter correction for NMR datasets. *Chemom. Intell. Lab. Syst. 131*, 1−6.