Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets

ARTICLE *in* ANALYTICAL CHEMISTRY · MAY 2006

Impact Factor: 5.64 · DOI: 10.1021/ac0519312 · Source: PubMed

Anal. Chem. **2006,** *78,* 2262−2267

Scaling and Normalization Effects in NMR Spectroscopic Metabonomic Data Sets

Andrew Craig, Olivier Cloarec, Elaine Holmes, Jeremy K. Nicholson, and John C. Lindon*

Biological Chemistry, Faculty of Natural Sciences, Imperial College London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ U.K.

Considerable confusion appears to exist in the metabonomics literature as to the real need for, and the role of, preprocessing the acquired spectroscopic data. A number of studies have presented various data manipulation approaches, some suggesting an optimum method. In metabonomics, data are usually presented as a table where each row relates to a given sample or analytical experiment and each column corresponds to a single measurement in that experiment, typically individual spectral peak intensities or metabolite concentrations. Here we suggest definitions for and discuss the operations usually termed normalization (a table row operation) and scaling (a table column operation) and demonstrate their need in ^{1}H NMR spectroscopic data sets derived from urine. The problems associated with "binned" data (i.e., values integrated over discrete spectral regions) are also discussed, and the particular biological context problems of analytical data on urine are highlighted. It is shown that care must be exercised in calculation of correlation coefficients for data sets where normalization to a constant sum is used. Analogous considerations will be needed for other biofluids, other analytical approaches (e.g., HPLC−MS), and indeed for other "omics" techniques (i.e., transcriptomics or proteomics) and for integrated studies with "fused" data sets. It is concluded that data preprocessing is context dependent and there can be no single method for general use.

The application of metabonomics has increased dramatically over the past few years, and many research groups are now attempting to process complex metabolic data sets. Metabonomics has been defined as the quantitative measurement of the multiparametric metabolic response of biological systems to pathology or genetic modification,[1] and the related subject of metabolomics has been described as a comprehensive analysis in which all the metabolites of a biological system are identified and quantified.[2] In these approaches, many samples from a biological origin (biofluids such as urine or plasma, tissue or plant extracts, in vitro culture supernatants, etc.) are analyzed using techniques that produce simultaneous detection and, in some cases, quantitation of many metabolites. The two main technologies that have been used so far have been ^{1}H NMR spectroscopy[3] and mass spectrometry with a prior on-line separation step such as HPLC[4] or GC (in this case, the metabolites generally have to be chemically derivatized).[5]

Studies to date have operated to a wide variety of experimental designs that are not always reported comprehensively. In addition, the data sets are usually preprocessed to make them amenable to multivariate statistical analysis. There have been a number of initiatives recently to build consensus on the standardization of metabolic experiments including the output from the SMRS group,[6,7] the ArMet scheme for reporting data and metadata,[8,9] and the output from the recent NIH-sponsored workshop.[10,11] Thus, considerable effort has been expended recently in order to define best practice in the field of metabonomics. These include comprehensive collection of background data to define the study samples (metadata), the techniques and their parameters used for data collection, and the chemometric and statistical procedures for data analysis.

The steps involved in the analysis of metabonomics data have been well described[6] and typically involve at a minimum: (a) postinstrument processing of acquired spectroscopic data, such as polynomial baseline correction by removal of offsets, calculation of intensity values either on each data point, on each peak, or summed over segmented regions (binning); (b) production of a data table from the analytical measurements such that there are m rows (observations, samples) and n columns (variables,

* To whom correspondence should be addressed. E-mail: j.lindon@imperial.ac.uk. Phone: +44 20 7594 3194. Fax: +44 20 7594 3066.

(1) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* **1999**, *29*, 1181−89.
(2) Fiehn, O. *Plant Mol. Biol.* **2002**, *48*, 155−71.
(3) Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Prog. NMR Spectrosc.* **2001**, *39*, 1−40.
(4) Wilson, I. D.; Plumb, R.; Granger, J.; Major, H.; Williams, R.; Lenz, E. *J. Chromatogr., B* **2005**, *817*, 67−76.
(5) Szopa, J.; Wilczynski, G.; Fiehn, O. *Phytochemistry* **2001**, *58*, 315−20.
(6) Lindon, J. C.; Nicholson, J. K.; Holmes, E.; Keun, H. C.; Craig, A.; Pearce, J. T.; 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.; Lehmann-Keeman, 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.; Spraul, M. *Nat. Biotechnol.* **2005**, *23*, 833−8.
(7) http://www.smrsgroup.org.
(8) Jenkins, H.; Hardy, N.; Beckmann, M.; Draper, J.; Smith, A. R.; Taylor, J.; Fiehn, O.; Goodacre, R.; Bino, R. J.; Hall, R.; Kopka, J.; Lane, G. A.; Lange, B. M.; Liu, J. R.; Mendes, P.; Nikolau, B. J.; Oliver, S. G.; Paton, N. W.; Rhee, S.; Roessner-Tunali, U.; Saito, K.; Smedsgaard, J.; Sumner, L. W.; Wang, T.; Walsh, S.; Wurtele, E. S.; Kell, K. B. *Nat. Biotechnol.* **2004**, *22*, 1601−6.
(9) http://www.armet.org.
(10) Castle, A. C.; Fiehn, O.; Kaddurah-Daouk, R.; Lindon, J. C. *Briefings Bioinformatics*. In press.
(11) http://www.niddk.nih.gov/fund/other/metabolomics2005.

10.1021/ac0519312 CCC: $33.50
Published on Web 02/18/2006

frequencies, integrals); (c) normalization of the data or some related adjustment to the spectral intensities (a row operation); (d) scaling of the data (a column operation); (e) multivariate statistical modeling of the data.

However, one major area where considerable interlaboratory variation exists is in the operating procedures used in stages c and d above to preprocess the spectroscopic data for chemometric analyses. One approach that has shown promise has been to adopt a pragmatic methodology for large-scale human studies.[12] An attempt has been made recently to define an optimum approach for NMR spectra of urine,[13] but the data analysis and results were not conclusive, and we show here that data preprocessing has to be context dependent. One particular approach that has been widely used is "binning" of spectra to produce a reduced set of parameters. This usually involves integration of peak values within defined specific spectral ranges, and the earlier attempt at defining an optimum approach used such binned data.[13]

The data preprocessing needs were also addressed by the SMRS group[6] and the NIH workshop,[10] but in the light of conflicting practices, it is thought timely to summarize best practice in this area and show that a single approach will not be optimum in all cases even for a given spectroscopic technique such as ^{1}H NMR spectroscopy. Thus, data for each study will have to be processed in an appropriate manner according to the study and type of sample. Here we use ^{1}H NMR spectroscopy as a vehicle for the discussions, but similar considerations will also apply for HPLC−MS metabonomic studies and in research involving other multivariate "omics" data, such as from transcriptomics and proteomics. The problems will be particularly acute for fused data sets such as combined NMR and MS data[14] or for data from combinations of different MS experiments.[15]

Given considerable confusion in the literature, we suggest here a standardized definition for, and explanation of several widely used and generally accepted practices.

Binning. Historically, pattern recognition of NMR-based metabolite data was performed using either quantitative or scored integrals of specific spectral peaks.[16,17] This approach does not work well in crowded regions of spectra with substantial peak overlap and is not easily automated for application to large sample sets. Calculating the peak areas within specified segments of a spectrum (binning or bucketing) was introduced originally to allow comparison of NMR data measured at different magnetic field strengths by minimizing but not eliminating the effects of second-order spectral differences.[18] In NMR-based metabonomic studies, it also reduces the effect of pH-induced changes in chemical shift, ensuring that the same species is always counted correctly across samples with such variation. Binning offers a rapid and consistent method by which to generate data sets automatically for modeling purposes, and a determination of the effects of changing the bin width and other variables for ^{1}H NMR spectra of biofluids such as urine and blood plasma or serum had been made previously.[18,19] A typical bin width of 0.04 ppm is frequently used for ^{1}H NMR spectra of urine as it is a good compromise between resolution and the difficulties related to positional variation in the position of some analyte species NMR peaks, most notably those of citrate, which are a function of pH and ionic strength of the sample. It also encompasses the typical width of an NMR resonance taking into account spin−spin splittings and line widths. By no means, is this bin width the only one that should be used.

Binned data should only be used for development of chemometric classification models, and it is necessary to examine and analyze the full resolution spectra for biomarker identification.

Since the use of binned data can lead to inaccuracies in peak intensities (e.g., by inclusion of variable amounts of baseline offset) and because of the removal of a number of computational limitations on data matrix sizes, more attention has recently been paid to methods that forego the need for binning of metabonomics data. These have involved methods of analyzing data in full resolution[20] both with and without additional preprocessing methods such as peak alignment.[21−23] Full spectral resolution data analysis is a significant advance allowing direct interpretation of any chemometric-derived model, since at full resolution these have a high similarity in appearance to real spectra, and this aids interpretability.

Normalization. This is a row operation that is applied to the data from each sample and comprises methods to make the data from all samples directly comparable with each other. A common use is to remove or minimize the effects of variable dilution of the samples. This is not usually a problem for samples from a biofluid such as plasma, where metabolite concentrations are highly regulated by homeostasis and changes observed in pathological situations are small but significant, or for studies on tissue extracts, where a constant or known weight of tissue can be analyzed. In such cases, it is desirable that the data should reflect directly the concentrations or relative concentrations of metabolites. In fused MS data sets[15] or in NMR spectra where each substance can give several peaks, it is conceivable that the same metabolite will be measured more than once and hence there is no a priori reason these data should be different. A major effect can be observed in NMR spectroscopic studies of urine where many drugs, toxins, and treatments can cause both large increases and decreases in urinary volume and hence in urinary concentration. Hence, for studies on urine, a good working definition requires an awareness of the biological context. Henceforth in this section, we confine ourselves to the discussion of

(12) Bijlsma, S.; Bobeldijk, I.; Verheij, E. R.; Ramaker, R.; Kochhar, S.; Macdonald, I. A.; van Ommen, B.; Smilde, A. K. *Anal. Chem.* **2006**, *78*, 567−74.

(13) Webb-Robertson, B. J.; Lowry, D. F.; Jarman, K. H.; Harbo, S. J.; Meng, Q. R.; Fuciarelli, A. F.; Pounds, J. G.; Lee, K. M. *J. Pharm. Biomed. Anal.* **2005**. *39*, 830−6.

(14) Crockford, D. J.; Holmes, E.; Lindon, J. C.; Plumb, R. S.; Zirah, S.; Bruce, S. J.; Rainville, P.; Stumpf, C. L.; Nicholson, J. K. *Anal. Chem.* **2006**, *78*, 363−371.

(15) Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van der Vat, B. J. C.; Jellema, R. H. *Anal. Chem.* **2005**, *77*, 6729−36.

(16) Gartland, K. P.; Sanins, S. M.; Nicholson, J. K.; Sweatman, B. C.; Beddell, C. R.; Lindon, J. C. *NMR Biomed.* **1990**. *3*, 166−72.

(17) Gartland, K. P.; Beddell, C. R.; Lindon, J. C.; Nicholson, J. K. *Mol. Pharmacol.* **1991**, *39*, 629−42.

(18) Spraul, M.; Neidig, P.; Klauck, U.; Kessler, P.; Holmes, E.; Nicholson, J. K.; Sweatman, B. C.; Salman, S. R.; Farrant, R. D.; Rahr, E.; et al. *J. Pharm. Biomed. Anal.* **1994**, *12*, 1215−25.

(19) Holmes, E.; Foxall, P. J.; Nicholson, J. K.; Neild, G. H.; Brown, S. M.; Beddell, C. R.; Sweatman, B. C.; Rahr, E.; Lindon, J. C.; Spraul, M.; et al. *Anal. Biochem.* **1994**, *220*, 284−96.

(20) Cloarec, O.; Dumas, M. E.; Trygg, J.; Craig, A.; Barton, R. H.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *Anal. Chem.* **2005**, *77*, 517−26.

(21) Stoyanova, R.; Nicholson, J. K.; Lindon, J. C.; Brown, T. R. *Anal. Chem.* **2004**, *76*, 3666−74.

(22) Stoyanova, R.; Nicholls, A. W.; Nicholson, J. K.; Lindon, J. C.; Brown, T. R. *J. Magn. Reson.* **2004**, *170*, 329−35.

(23) Forshed, J.; Torgrip, R. J.; Aberg, K. M.; Karlberg, B.; Lindberg, J.; Jacobsson, S. P. *J. Pharm. Biomed. Anal.* **2005**, *38*, 824−32.

normalization of data derived from NMR-based metabonomic analysis of urine.

The variable nature of urinary volume is often overlooked in metabonomic studies of urine.[13] Renal excretion and urine production serves two purposes, to eliminate waste organic and inorganic species and also to help regulate blood volume. This can be conceived as being achieved through independent control of both solute and solvent excretion rates. The need for normalization in NMR-based metabonomics studies of urine arises primarily from this biological imperative. To correctly obtain true values for the excretion of urinary metabolites, it is necessary to measure the absolute values in molar concentration units per unit time per unit of body weight (e.g., millimoles, per hour, per kilogram of body weight). In animal studies, this of course requires measurement of urinary volumes and animal weights as a study proceeds. For clinical studies, the same criteria also apply. In practice, this method is almost never fully employed. However, one procedure that can help to experimentally normalize such NMR data is to note the volume of urine from which the sample came, freeze-dry the urine, and then reconstitute it in an appropriate volume of water or buffer to account for any urinary volume differences for a given collection period.[24] Of course, depending on the context of the problem, in many studies, it is perfectly possible to use relative concentrations of metabolites and changes of these as biomarkers of a biological effect. For studies of excretion balance or for flux calculations, often absolute metabolite concentrations and amounts are required, but for investigations of the differential effects of drugs or toxins, for example, it might be that relative effects are sufficient.

Normalization to a "housekeeping" metabolite has also been attempted, typically using a creatinine peak area as a reference.[25] Creatinine clearance has long been considered a constant and used to assess renal function though this may not always be the case as, in some cases, creatinine clearance may be more related to muscle mass particularly in children[25] and the elderly.[26] This results in a metabolite excretion rate relative to that of creatinine. Alternatively, in some cases, the concentration of a specific metabolite can be determined by an independent means (e.g., glucose using conventional clinical chemistry) and this then provides a reference value.[27]

It is more common, however, for other preprocessing methods to be used before chemometric modeling, where the biological meaning of the procedure is not directly apparent. One common method of normalization involves setting each observation (spectrum) to have unit total intensity by expressing each data point as a fraction of the total spectral integral. We refer to this method as normalization to a constant sum (CS).

It might be useful to consider the volume of urine produced by an individual in a given time (V_{total}) as being composed of two parts. The first is the volume of water required to dissolve all excreted solutes (V_{solute}). The second is the volume ($V_{balance}$), which could be positive or negative, that serves to balance fluid intake in order to maintain blood volume. In principle, it would be best to compare the metabolite concentrations in V_{solute} between subjects, but this is not possible since only V_{total} is known. The value of $V_{balance}$ will vary between individuals and will be affected by the particular study in progress. For a series of spectra with highly similar internal peak ratios but differing in total intensity because of such dilution or concentration effects, CS normalization of each spectrum can be considered to approximate the relative concentration of species (i.e., as in V_{solute}). Importantly, this approximation will break down when large perturbations occur to intensities in some spectra (e.g., those from certain toxin-treated animals or the use of diuretic drugs, for example). This is easily seen because if some peak areas are increased and the total is normalized to a constant, others will *appear* to have decreased, and this effect in "closed" data sets has been noted previously.[28] This can cause a major difficulty in the interpretation of loadings from chemometric analyses such as principal components analysis (PCA) or PLS when such CS normalization has been used.

Scaling. This operation is performed on the columns of data (i.e., on each spectral intensity across all samples). A number of such scaling methods are commonly used. Each column of the table can be given a mean of zero by subtracting the column mean from each value in the column (mean-centering). This is typically done so that all the components found by PCA have as their origin the centroid of the data, resulting in a parsimonious model. Second, each column of the table can be scaled so that it has unit variance, by dividing each value in the column by the standard deviation of the column. If the data are mean centered, the weighting reflects the covariance of the variables, while in unit variance scaling, the weighting reflects their correlation. Other forms of scaling are possible[29] and are used, such as Pareto scaling, where each variable is divided by the square root of the standard deviation of the column values,[30] or logarithmic scaling, when values relating to order of magnitude scores are desired.[27]

Furthermore, it should be clear that as normalization and scaling operations serve different purposes it is possible (and is, in fact, the usual practice) to use a combination of normalization and scaling methods.

RESULTS

To demonstrate the consequences of normalization to a constant sum and column scaling, a set of simulated NMR spectra were constructed and visualized using Matlab routines and analyzed using PCA. In this example, each of 40 samples contains two components, each of which results in a single Lorentzian-shaped peak at NMR chemical shifts of 3 and 7 ppm, with a full width at half-height of 0.1 ppm set in a spectral range of 1−10 ppm, digitized in 0.01 ppm steps (Figure 1A). The samples are divided into two classes (such as control and diseased) with peak 1 at 3 ppm with a variance of 10 and a mean value of 100 in class 1 and a mean value of 120 in class 2. Peak 2 at 7 ppm is not class

(24) Gartland, K. P.; Bonner, F. W.; Nicholson, J. K. *Mol. Pharmacol.* **1989**, *35*, 242−50.

(25) Prevot, A.; Martini, S.; Guignard, J. P. *Rev. Med. Suisse Romande* **2002**, *122*, 625−30.

(26) Burkhardt, H.; Bojarsky; Gretz, N.; Gladisch, R. *Gerontology* **2002**, *48*, 140−6.

(27) Anthony, M. L.; Sweatman, B. C.; Beddell, C. R.; Lindon, J. C.; Nicholson, J. K. *Mol. Pharmacol.* **1994**, *46*, 199−211.

(28) Wold, S.; Johansson, E.; Cocchi, M. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Perspectives in Drug Discovery and Design; ESCOM Science: Leiden, 1998.

(29) Holmes, E.; Nicholls, A. W.; Lindon, J. C.; Ramos, S.; Spraul, M.; Neidig, P.; Connor, S. C.; Connelly, J.; Damment, S. J.; Haselden, J.; Nicholson, J. K. *NMR Biomed.* **1998**, *11*, 235−44.

(30) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis. Principles and Applications*; Umetrics AB: Umeå, Sweden, 2001.

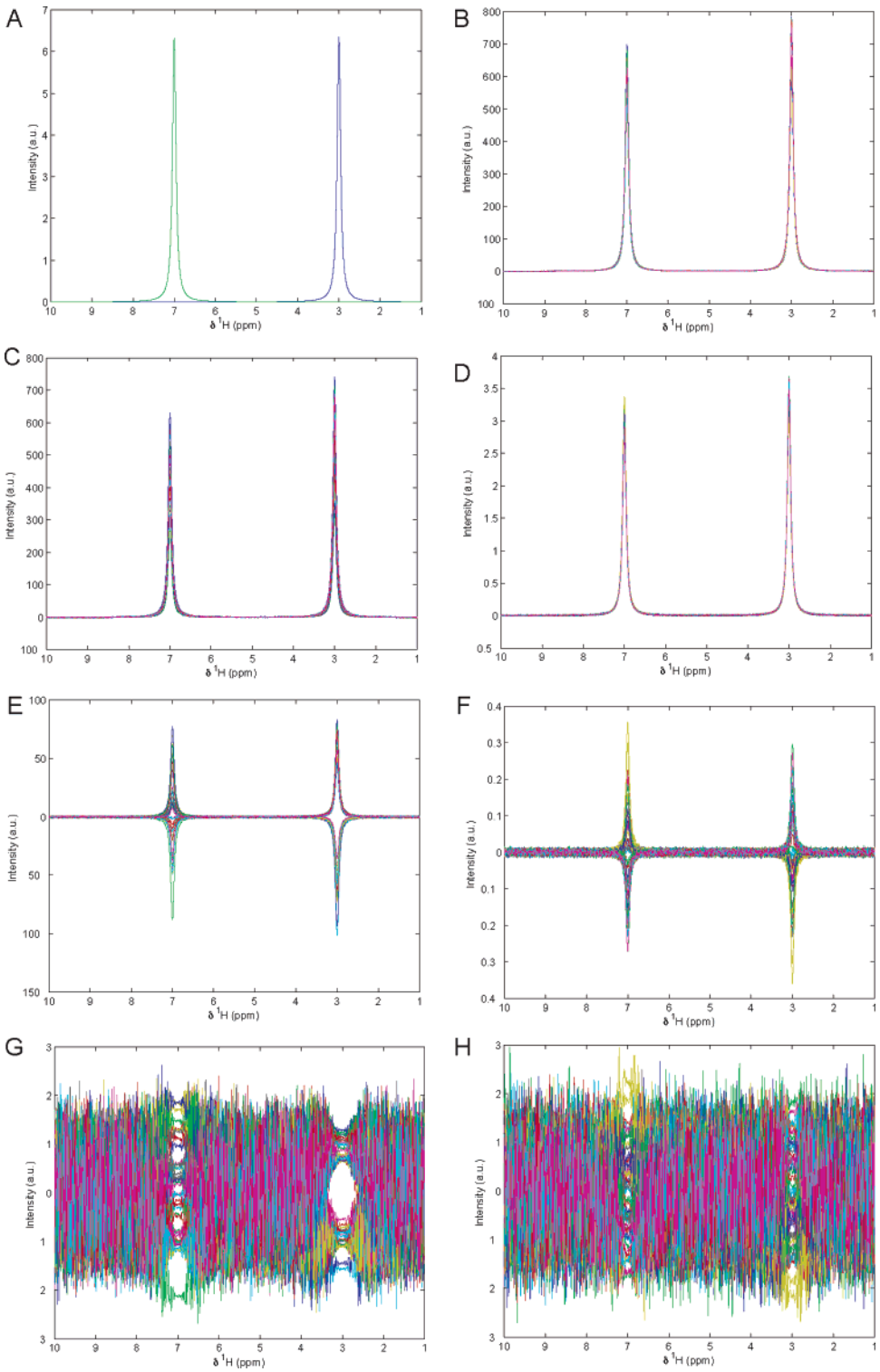

Figure 1. Simulated spectrum comprising two Lorentzian peaks. (A) The two peaks before the additional of noise, (B) superposition of all 40 spectra, (C) superposition of all 40 spectra, with random dilutions, (D) same as (C) but all spectra normalized to a constant sum, (E) the mean-centered true data set, (F) the mean-centered normalized data set, (G) the mean-centered and unit variance scaled true data set, and (H) the mean-centered and unit variance scaled normalized data set.

discriminating and has an intensity that is normally distributed across all observations with a mean of 100 and a variance of 50.

The true data matrix, **X** (Figure 1B), was constructed as $\mathbf{X} = \mathbf{C}\cdot\mathbf{P} + \mathbf{E}$, where **C** is a concentration matrix, **P** is a matrix of pure compound spectra, and **E** is a matrix of white noise with values between ±1 representing thermal noise from the spectrometer. A "diluted" version of the data matrix **X** (Figure 1C) was also produced where $\mathbf{X}_d = \mathbf{D}\cdot\mathbf{C}\cdot\mathbf{P} + \mathbf{E}$, and **D** is a diagonal matrix containing random dilution factors between 0.5 and 1.0 to represent a variation in urinary volumes that is independent of any class differences. The diluted data set was normalized using the CS method, and the result is shown in Figure 1D. Because

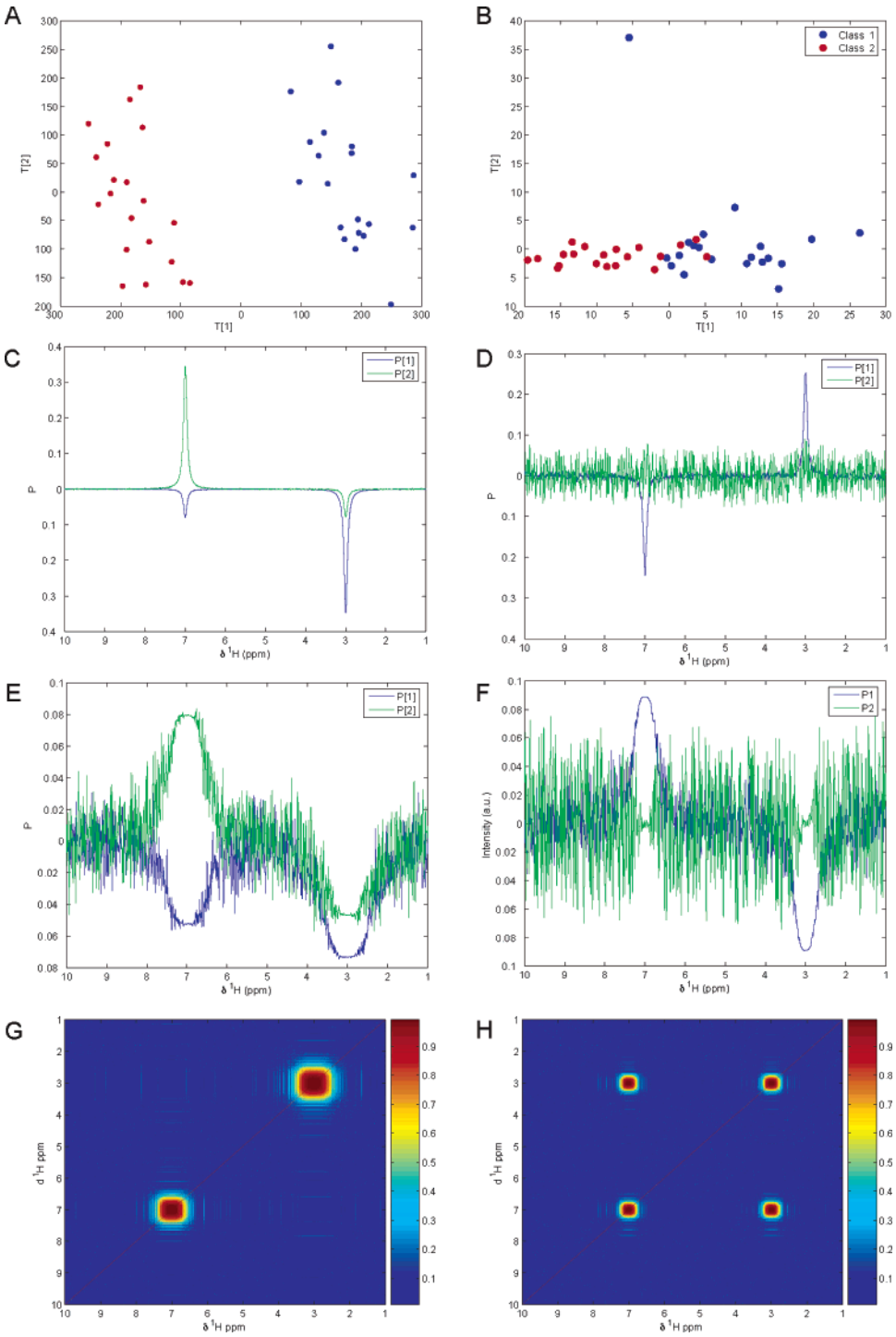

Figure 2. Results of PCA on the simulated spectral data sets. (A) Scores plot PC1 vs PC2 for the mean-centered true data, (B) scores plot PC1 vs PC2 for the mean-centered normalized data, (C) PC1 vs PC2 loadings plot for the mean-centered true data, (D) PC1 vs PC2 loadings plot for the mean-centered normalized data, (E) PC1 vs PC2 loadings plot for the mean-centered and unit variance scaled true data, (F) PC1 vs PC2 loadings plot for the mean-centered and unit variance scaled normalized data, (G) STOCSY correlation plot for the true data, and (H) STOCSY correlation plot for the normalized data.

the instrumental noise has a constant average value, when very dilute spectra are normalized to a constant sum, the apparent signal-to-noise ratio is decreased. The effect of mean-centering the variables is shown on the true data (Figure 1E) and on the normalized data (Figure 1F). The effect of scaling each variable to have unit variance is shown on both data sets (Figure 1G and H, respectively).

Both data sets were then modeled using PCA. The expected result is a two-component model showing the two classes clearly separated in PC1 with the corresponding PC1 loadings identifying the pure compound spectrum (i.e., peak 1) of the species responsible for the difference between the two classes. Indeed, for the PC scores for mean-centered true data, as shown in Figure 2A, this is the result obtained. The unit variance scaled true data

gave a model with a similar result (not shown). The corresponding loadings are shown for the mean-centered true data (Figure 2C) and for the unit variance scaled true data (Figure 2E). In both cases, the loadings for PC1 identify peak 1 as the feature that describes the variance causing the separation between the two classes and the PC2 loadings describe the variation in peak 2 intensity.

Mean-centering results in a model with loadings that have a pseudospectral appearance retaining Lorentzian line shapes. The use of unit variance scaling gives equal weight to each data value, allowing systematic changes with small variance to be more readily detected. However, when the loadings are plotted as a pseudospectrum, the apparent signal-to-noise ratio is degraded and there is a tendency toward square-shaped peaks since intensity values within the same peak will be highly correlated and will thus receive similar weights in the PCA. In the case of NMR data, this can confound the useful information obtainable on peak height ratios and peak multiplicities. The recently reported back-scaling method for some chemometrics procedures (O-PLS) attempts to combine the advantages of both of these approaches.[20]

The scores for the normalized data are given in Figure 2B, and again, the first PC largely discriminates between the two classes, but an outlier can be seen in the second PC. This is explained as arising from a very dilute sample with a low signal-to-noise ratio, since PC2 is only modeling noise in the data. The corresponding loadings are shown for the mean-centered model (Figure 2D) and for the unit variance scaled model (Figure 2F). After normalization, in both cases, the loadings indicate that it is apparently a combination of an increase in peak1 intensity and a decrease in peak 2 intensity that is the discriminating feature. This is because the CS normalization has mixed the variation in one peak in with that from the other peak and changed the correlation structure of the data. Hence, the real discriminating factor in the data set has been obscured. This change in correlation structure is readily apparent when the STOCSY technique[31] is used, as shown in Figures 2G for the true data set and Figure 2H for the normalized set. STOCSY, as expected, does not show any cross-peak indicating covariance for the true data set, but a cross-peak is observed for the normalized data set. Artifactual cross-peaks that arise as a result of the choice of preprocessing method used may thus confound the use of STOCSY for structural interpretation, if used inappropriately.

CONCLUSIONS

We acknowledge that there is little exploration of normalization of intersample volume-related variation within the metabonomic and metabolomic literature. In many cases, exploratory data analysis is carried out by varying normalization and scaling procedures in order to obtain an optimum separation of two or more sample classes using chemometrics methods such as PCA, PLS, or neural networks. While this is useful for deriving classification models for predicting the class of subsequent samples, the interpretation of the biochemical factors responsible for the classification is not straightforward.

Thus, we arrive at a more appropriate definition such that normalization of urinary metabolic data should be best considered as a data transformation which minimizes intersample variation due to differences in gross urinary concentration between samples caused by volume differences. The effect of CS normalization as demonstrated here primarily affects the interpretation of chemometric model coefficients. The CS normalization method has been used to good effect to produce models and expert systems with good predictive power.[29]

Methods analogous to the invariant set normalization for microarray data[32] as well as scaling by the gradient of a robust regression to the median sample[33] might yield some success as alternative preprocessing methods in metabonomics. It should also be clear that any normalization should be a linear transformation.

Finally, with the increased interest in relating data sets collected from different analytical platforms or across different levels of molecular biology, it is important to be aware of the effects of preprocessing on statistical outcome and to be mindful of the consequences of a chosen method of preprocessing and the limitations that this will place on the interpretation of any chemometric model.

ACKNOWLEDGMENT

We thank the Wellcome Trust for financial support (to O.C. and A.C.) from the project, Biological Atlas of Insulin Resistance (www.bair.org.uk).

Received for review October 28, 2005. Accepted January 13, 2006.

AC0519312

(31) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. *Anal. Chem.* **2005**, *77*, 1282−9.

(32) Tseng, G. C.; Oh, M. K.; Rohlin, L.; Liao, J. C.; Wong, W. H. *Nucleic Acids Res.* **2001**, *29*, 2549−57.

(33) Craig, A. Parallel metabonomic and genomic characterisation of experimental hepatotoxicity in the rat. Ph.D. Thesis, University of London, 2004; p 293.