Statistical Total Correlation Spectroscopy Scaling for Enhancement of Metabolic Information Recovery in Biological NMR Spectra

ARTICLE *in* ANALYTICAL CHEMISTRY · DECEMBER 2011

Impact Factor: 5.64 · DOI: 10.1021/ac202720f · Source: PubMed

Judith Marlou Fonville
University of Cambridge
19 PUBLICATIONS **405** CITATIONS
SEE PROFILE

Muireann Coen
Imperial College London
54 PUBLICATIONS **1,648** CITATIONS
SEE PROFILE

Caroline Rae
University of New South Wales
117 PUBLICATIONS **2,690** CITATIONS
SEE PROFILE

Jeremy K Nicholson
Imperial College London
754 PUBLICATIONS **43,253** CITATIONS
SEE PROFILE

analytical chemistry

Article

pubs.acs.org/ac

Statistical Total Correlation Spectroscopy Scaling for Enhancement of Metabolic Information Recovery in Biological NMR Spectra

Anthony D. Maher,*,[†,‡,§] Judith M. Fonville,[§] Muireann Coen,[§] John C. Lindon,[§] Caroline D. Rae,[†] and Jeremy K. Nicholson*,[§]

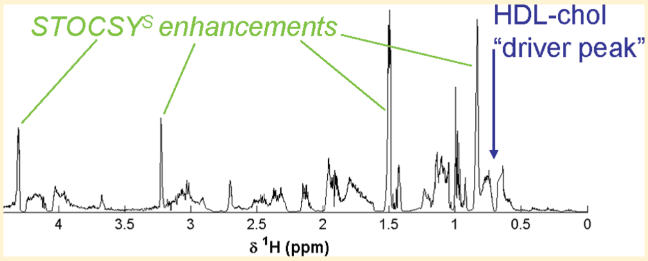
[†]Neuroscience Research Australia, Barker Street, Randwick 2031, Australia

[‡]School of Medical Sciences, University of New South Wales, New South Wales 2052, Australia

[§]Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, SW7 2AZ London, United Kingdom

Supporting Information

ABSTRACT: The high level of complexity in nuclear magnetic resonance (NMR) metabolic spectroscopic data sets has fueled the development of experimental and mathematical techniques that enhance latent biomarker recovery and improve model interpretability. We previously showed that statistical total correlation spectroscopy (STOCSY) can be used to *edit* NMR spectra to remove drug metabolite signatures that obscure metabolic variation of diagnostic interest. Here, we extend this "STOCSY editing" concept to a generalized scaling procedure for NMR data that enhances recovery of latent biochemical information and improves biological classification and interpretation. We call this new procedure STOCSY-scaling (STOCSY^S). STOCSY^S exploits the fixed proportionality in a set of NMR spectra between resonances from the same molecule to suppress or enhance features correlated with a resonance of interest. We demonstrate this new approach using two exemplar data sets: (a) a streptozotocin rat model (n = 30) of type 1 diabetes and (b) a human epidemiological study utilizing plasma NMR spectra of patients with metabolic syndrome (n = 67). In both cases significant biomarker discovery improvement was observed by using STOCSY^S: the approach successfully suppressed interfering NMR signals from glucose and lactate that otherwise dominate the variation in the streptozotocin study, which then allowed recovery of biomarkers such as glycine, which were otherwise obscured. In the metabolic syndrome study, we used STOCSY^S to enhance variation from the high-density lipoprotein cholesterol peak, improving the prediction of individuals with metabolic syndrome from controls in orthogonal projections to latent structures discriminant analysis models and facilitating the biological interpretation of the results. Thus, STOCSY^S is a versatile technique that is applicable in any situation in which variation, either biological or otherwise, dominates a data set at the expense of more interesting or important features. This approach is generally appropriate for many types of NMR-based complex mixture analyses and hence for wider applications in bioanalytical science.



Metabonomics is concerned with the measurement of the global metabolic responses of organisms to stimuli, such as a disease or a toxic episode.[1] Biofluids such as urine or blood plasma are a valuable source of information reflecting the overall biological state of an individual. Analytical technologies such as NMR and mass spectrometry (MS) can be used to measure many metabolites simultaneously in such biofluids.[2] The resulting metabolic "fingerprint" can then be subjected to statistical pattern recognition techniques to compare samples and identify characteristic metabolic variations specific to physiological processes. Commonly used pattern recognition routines for NMR data are linear approaches such as principal component analysis (PCA) and projections to latent structures (PLS),[3] but other methods are being increasingly used.[4,5]

A common feature of metabonomic data is the high level of spectral complexity. This exists because of the large number of metabolites measured simultaneously, compounded by the multiple sources of variation that can influence metabolic profiles, whether biological, environmental, or technical. The nature of NMR also adds to this complexity, with many metabolites showing multiple resonances and with some resonances subject to peak overlap and positional frequency variation. Therefore, a number of methods have been proposed to simplify the data at acquisition (by NMR experimental means) or postacquisition (e.g., by mathematical means). Examples of experimental simplification include applying a Carr−Purcell−Meiboom−Gill (CPMG)[6] spin-echo pulse sequence, which selectively suppresses signals from larger molecules with shorter T_2 relaxation times, and the use of J-resolved pulse sequences to remove overlap in crowded spectral regions.[7] Alternatively, data may be simplified after acquisition:

Received: October 14, 2011
Accepted: December 8, 2011
Published: December 8, 2011

a “virtual” CPMG experiment may be generated from a 1D data set by mathematical manipulation,[8] and peak alignment can be used to remove positional shifts.[9]

Data can also be modified to ease information extraction through scaling, a mathematical modification of the variables to adjust the original distribution of variance in the data. The type of scaling applied to a data set is important since it can influence the results of statistical analyses.[10] Unscaled data typically over-represent high-intensity peaks, and smaller, possibly more important, signals are mostly disregarded. This imbalance can be overcome by a type of scaling commonly applied to ^{1}H NMR data: unit-variance (UV) scaling, where the values for each variable are mean centered by subtraction of the average value and then divided by the variable standard deviation. The drawback of UV-scaling is that spectral regions containing noise have variance equal to that of regions containing “real” signals. Interpretation and biomarker identification are then hampered because the loading plots of multivariate models no longer resemble the original spectra. Pareto scaling can be used as a compromise between no scaling and UV-scaling, where the variable is instead divided by the square root of its standard deviation. It has been noted that inclusion and exclusion of variables can also be considered a form of scaling.[11] For example, exclusion of the variables containing the residual water peak and urea in ^{1}H NMR spectra from urine is routine practice in metabonomics.[12] This is equivalent to multiplying these variables by 0, while all others are multiplied by 1.

An alternative approach to improve model interpretation and reduce data set complexity, in addition to the widely applied dimensionality reduction techniques such as PCA and PLS and various deconvolution techniques, is orthogonal signal correction.[13−15] This methodology removes the variation in a data set that is orthogonal (i.e., unrelated) to the response variable of interest. When combined with projection to latent structures, then called orthogonal projections to latent structures (OPLS), it is a means of removing variation from a matrix called **X** (e.g., the intensity values in the digitized spectra), not correlated with a response matrix, **Y**. This facilitates the interpretation of these chemometric models, while retaining the same predictive ability as PLS, which is especially useful if the source of confounding variation is unknown. However, there are often cases in metabonomic studies where the confounding variation in NMR data is well-defined. A common example is the presence of peaks from drug metabolites in urine spectra in toxicology studies, which can obscure more subtle endogenous changes that result from toxicity.[16] Another example is the appearance of glucose peaks in biofluids from patients with diabetes.[17] Although glucose is endogenous, this marked response is already well characterized in diabetes, and NMR peaks of glucose often obscure other peaks in the same chemical shift region, which is exacerbated by the fact that the NMR spectrum of glucose has many peaks; it can be considered to occupy a large chemical shift range in metabonomics data sets from blood plasma.[7]

Statistical spectroscopic methods such as statistical total correlation spectroscopy (STOCSY) have proven valuable tools for information recovery in NMR data sets and work by exploiting inherent colinearities in sets of NMR spectra acquired under comparable conditions.[18−20] For example, we have shown STOCSY is an effective means of detecting xenobiotic drug metabolites in urine,[21] recovering biomarker information from heteronuclear NMR data sets[22,23] and integrating MS and NMR data for enhanced metabolite assignment.[24] Several other groups have also adapted this technology for extraction of biological information from NMR data sets from complex mixtures.[25−27]

We have recently shown that STOCSY can be adapted to *edit* NMR spectra to enhance information recovery from toxicological data (STOCSY-E).[16] This procedure has recently been further developed to a more in-depth, iterative process to recover information from toxicology data sets (STOCSY-I).[28] In this paper, we introduce a new STOCSY-based scaling procedure that reveals subtle changes relevant to the biology of the samples that may otherwise have been obscured. Unlike previous work, this new approach does not require any defined “cutoff” value for extraction of metabolic information from NMR spectral data. We call this new approach STOCSY-scaling (STOCSY^S).

■ MATERIALS AND METHODS

Materials. The chemicals used in this study were purchased from Sigma (St. Louis, MO). D_2O (99.9%) was from Goss Scientific Instruments Ltd. (Essex, U.K.).

Streptozotocin Study. As part of the COMET project,[29] 7 week old male Sprague−Dawley rats (n = 30) were dosed with streptozotocin (in 10 mM citrate buffer) at 0, 25, and 60 mg kg^{-1}. Blood samples were collected at 48 and 168 h postdose, allowed to clot, and centrifuged. The serum was immediately frozen at −40 °C. Thawed samples were diluted 2:1 in 0.9% (w/v) NaCl in 20% D_2O. ^{1}H NMR data from plasma samples were acquired on a Bruker Avance spectrometer operating at 600.29 MHz using a 5 mm TXI flow injection probe equipped with a z-gradient (Bruker, Rheinstetten, Germany) at 300 K. The CPMG pulse sequence d_1−90°−$(\tau/2$−180°−$\tau/2)_n$−acquire was applied for 128 transients with 8 dummy scans and a total echo time of 64 ms (n = 80, $\tau/2$ = 400 μs).[6,30] The spectral width was 12 019 Hz, 32k time domain data points were acquired, and selective irradiation of the water resonance was applied during the relaxation delay, d_1. The data were phase and baseline corrected using in-house software (NMRPROC, T. M. D. Ebbels and H. C. Keun, private communication). The following spectral regions were determined by visual inspection for removal of glucose and lactate signals: δ 1.316−1.345 and δ 4.089−4.140 (lactate); δ 3.217−3.277, δ 3.380−3.693, and δ 4.613−5.269 (glucose).

Metabolic Syndrome Study. As part of the MolPAGE project (www.molpage.org), blood plasma was collected from a cohort of humans (the MolOBB cohort, n = 67) with or without metabolic syndrome (MetS) and immediately frozen at −40 °C. Thawed samples were centrifuged at 16060g for 5 min and diluted 1:4 in physiological saline (0.9% (w/v) NaCl) in 20% D_2O. A 1D ^{1}H NMR pulse sequence of the form d_1−90°−3 μs−90°−t_m−90°−acquire was used to obtain spectra, where selective irradiation of the water signal was applied during d_1 (2 s) and the mixing time, t_m (100 ms). A total of 96 transients were acquired over 64k data points with a sweep width of 12 019 Hz. Spectra were phase and baseline corrected using Topspin (Bruker).

Data Analysis. Prior to pattern recognition, ^{1}H NMR spectra were zero-filled once and Fourier transformed with 1 Hz exponential multiplication. Data were imported into MATLAB (MathWorks, Natick, MA) and referenced such that the anomeric proton resonance from α-glucose was set to δ 5.233. The spectral regions containing the residual water peak were removed. PCA and OPLS models were constructed using scripts written in-house based on published algorithms.[13,31] Orthogonal projections to latent structures discriminant analysis (OPLS-DA) models are a class of OPLS models where the response matrix consists of dummy variables (0s and 1s) corresponding to the class of each sample. The parameter Q^2Y was computed for all the OPLS models by 7-fold cross-validation. All models were mean centered after scaling.

RESULTS AND DISCUSSION

Implementation of STOCSY^S. The ^{1}H NMR spectrum from a blood plasma sample from a rat from the control (i.e., vehicle-dosed) group from the streptozotocin toxicity study, expanded to show the spectral region containing glucose resonances, is shown in Figure 1 A. Some peaks are "easier" to assign than others; for example, the relatively isolated anomeric α-glucose proton doublet at δ 5.23 and the lactate CH quartet at δ 4.11 are readily identified, whereas peaks from other metabolites are more difficult to assign because they are of lower intensity or overlap with those from glucose.

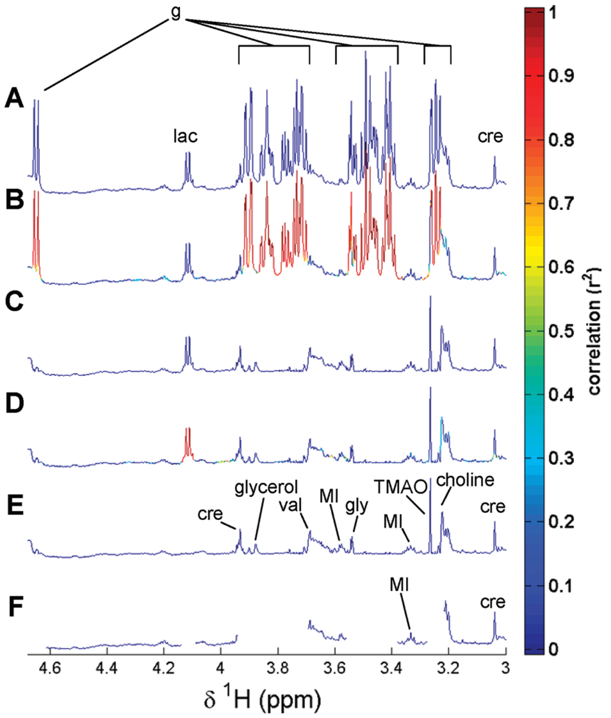


Figure 1. STOCSY^S to suppress glucose and lactate signals in the streptozotocin study. (A) ^{1}H NMR spectrum from a control rat plasma sample, expanded around the region between δ 3.0 and δ 4.7. The resonances that can be readily assigned are from glucose (g), lactate (lac), and creatine (cre). (B) Vector of correlation coefficients (r^2) from the glucose resonance at δ 5.23 plotted as a color onto the ^{1}H NMR spectrum. (C) Same spectrum as in (A), but after STOCSY^S of the glucose peaks. (D) Vector of correlations (r^2) from the lactate peak at δ 4.11 plotted as a color onto the scaled NMR spectrum from (C) used for the second round of STOCSY^S. (E) Same spectrum as in (C) after two rounds of STOCSY^S to remove first the glucose and then the lactate resonances. Additional metabolites are now identifiable and have been labeled: valine (val), trimethylamine *N*-oxide (TMAO), cre, glycine (gly), *myo*-inositol (MI). (F) Same spectrum as in (A) after excision of the regions containing glucose and lactate by the eye as described in the Materials and Methods.

Glucose gives rise to many highly coupled multiplets in the NMR spectrum and tends to dominate the variation in NMR data sets from blood or urine samples from patients with overt diabetes.[17] Thus, the removal of variables corresponding to glucose resonances has proven useful, for example, in investigations of the effect of streptozotocin on metabolic profiles of blood plasma and urine in rats[32] and the effects of thiazolidinediones in patients with type 2 diabetes,[33] where glucose resonances dominated the data set. However, removal of interfering variables is not a trivial task since the limits of an NMR peak are ambiguous; theoretically, an NMR peak line shape is described by a Lorentzian function that approaches zero only as the distance from the mean resonance frequency approaches infinity. Furthermore, in the case of peak overlap, it is not straightforward to determine whether all spectral regions known to contain glucose should be removed or whether some overlapping peaks could remain.

We propose to reduce the intensities of glucose-related resonances using the suggested STOCSY^S approach. STOCSY results in a vector of correlation coefficients, $\mathbf{r}_i$, where i runs over all data points in the spectrum;[18] an example of STOCSY "driven" from the glucose resonance at δ 5.23 is shown in Figure 1B. The color scale indicates the square of the correlation coefficient, and it is clear that all other glucose resonances are highly correlated to the anomeric α-glucose signal. In STOCSY^S, a known, relatively isolated peak (here from glucose) is selected and the spectra $\mathbf{X}_{ij}$ are edited on the basis of the calculated STOCSY correlations via eq 1, and the

$$\mathbf{X}_{Sij} = \mathbf{X}_{ij}(1 - \mathbf{r}_i^{\,2}) \qquad (1)$$

resulting spectrum is shown in Figure 1 C. A second round of STOCSY was then performed using the lactate resonance at δ 4.11; see Figure 1 D. This was done because high lactate levels in two samples were dominating the variation in the data set, without revealing any systematic information relating to the streptozotocin treatment. In the resultant spectrum, shown in Figure 1 E, the spectral regions containing glucose and lactate are reduced to nearly 0, and the non-glucose resonances in this region are now easier to resolve and assign. Peaks that were uncorrelated and nonoverlapped with glucose are unaffected, such as creatine at δ 3.04 and the *myo*-inositol pseudotriplet at δ 3.28. Resonances from other metabolites that were difficult to assign in the original spectrum are now more easily resolved such as a singlet at δ 3.26. This is from trimethylamine *N*-oxide (TMAO), which appears as a shoulder on the H2 resonance from β-glucose at δ 3.24. This exemplifies the advantage of the STOCSY^S approach: high- and low-frequency limits for where a peak begins and ends do not have to be defined, unlike the situation where resonances are simply "cut" from the spectrum, as was done to produce Figure 1 F. This also highlights a key difference between the STOCSY^S approach and our previous STOCSY-E method. In STOCSY^S no threshold value needs to be defined to discriminate intra- from intermolecular statistical connectivities, as each spectrum is simply scaled according to eq 1.[34] The advantage of using STOCSY^S over STOCSY-E is illustrated in Figure S1A, Supporting Information. This figure shows the results of the PCA analysis from the same data after the glucose and lactate resonances were edited out using the STOCSY-E procedure. Here a threshold (θ) of 0.9 was used to remove glucose and lactate. However, the overlap of the methyl lactate resonances with the lipoprotein signals resulted in poor attenuation at this threshold, and hence, the lactate signals dominate principal component 2 (PC2).

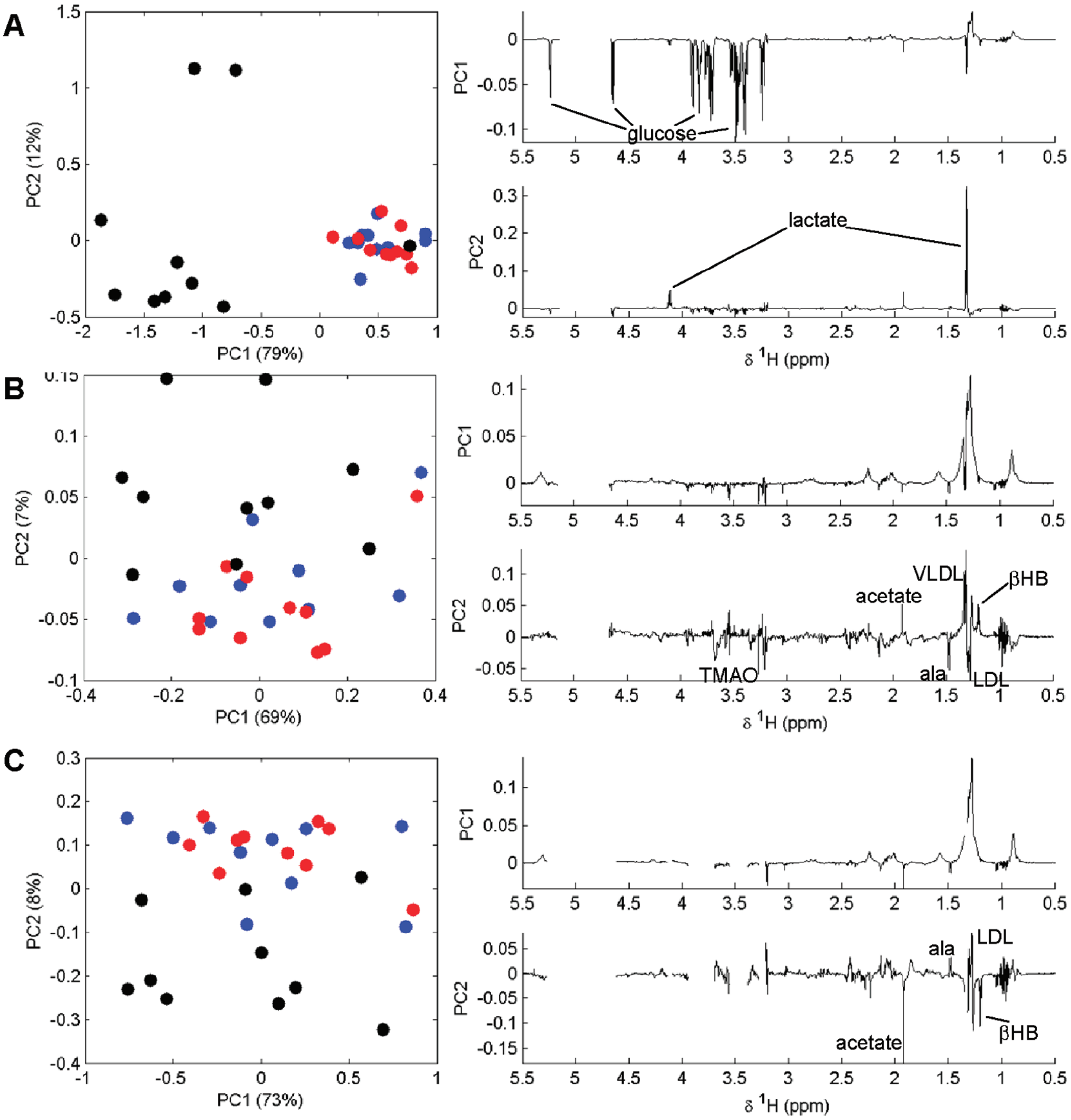


Figure 2. Scores (left panels) and loadings (right panels) plots from PCA models of the streptozotocin data set. PCA scores have been color-coded according to class: high dose (60 mg kg^{-1}), black; low dose (25 mg kg^{-1}), red; controls (vehicle), blue. Numbers in parentheses represent the percentage of variation explained in the respective PC. (A) PCA model based on the full data set. (B) PCA model after the glucose and lactate resonances were edited with STOCSY^{S}. (C) PCA model after manual removal of glucose and lactate resonances. TMAO = trimethylamine *N*-oxide, ala = alanine, LDL = low-density lipoprotein, and *β*HB = *β*-hydroxybutyrate.

It is also possible to compare STOCSY^{S} to OPLS-DA as a method for removing unwanted variation. For example, if the response matrix is constructed as the vector of intensities of the glucose resonance at *δ* 5.233, we can analyze the variation in the orthogonal component by PCA. The results of this are displayed in Figure S1B, Supporting Information, showing suppression of glucose resonances in the PC loadings and scores distribution dominated by lactate. However, when we attempt to remove two orthogonal components after constructing **Y** from both the glucose and lactate intensities, the glucose resonances are still dominant in the data set (Figure S1C).

Multivariate Analysis of STOCSY^{S} Data. Figure 2 A shows the scores and loadings plots for the PCA model derived from the streptozotocin serum data. With the exception of one nonresponder, the high-dose group (black) is completely separated from the low-dose group (red) and controls (blue) in principal component 1 (PC1). The loadings revealed that this was due to increased amounts of glucose in the high-dose group. Variation in PC2 was dominated by lactate resonances due to two samples in the high-dose group having high lactate levels. Although a good separation was achieved along PC1, the only readily made observation is that glucose is increased in rats treated with high doses of streptozotocin.

Figure 2 B shows the scores and loadings plots of PCA on the data after removal of glucose and lactate resonances according to the STOCSY^{S} procedure described above. With the exception of the nonresponder, the high-dose group separates out from the low-dose group and controls along the second principal component. The metabolites responsible for this separation can be established from the loadings plots as higher very-low-density lipoprotein (VLDL), acetate, and *β*-hydroxybutyrate (*β*HB) resonances and lower TMAO, choline, alanine, and low-density lipoprotein (LDL). These are consistent with previous observations from a metabonomic study on streptozotocin dosing in rats.[32]

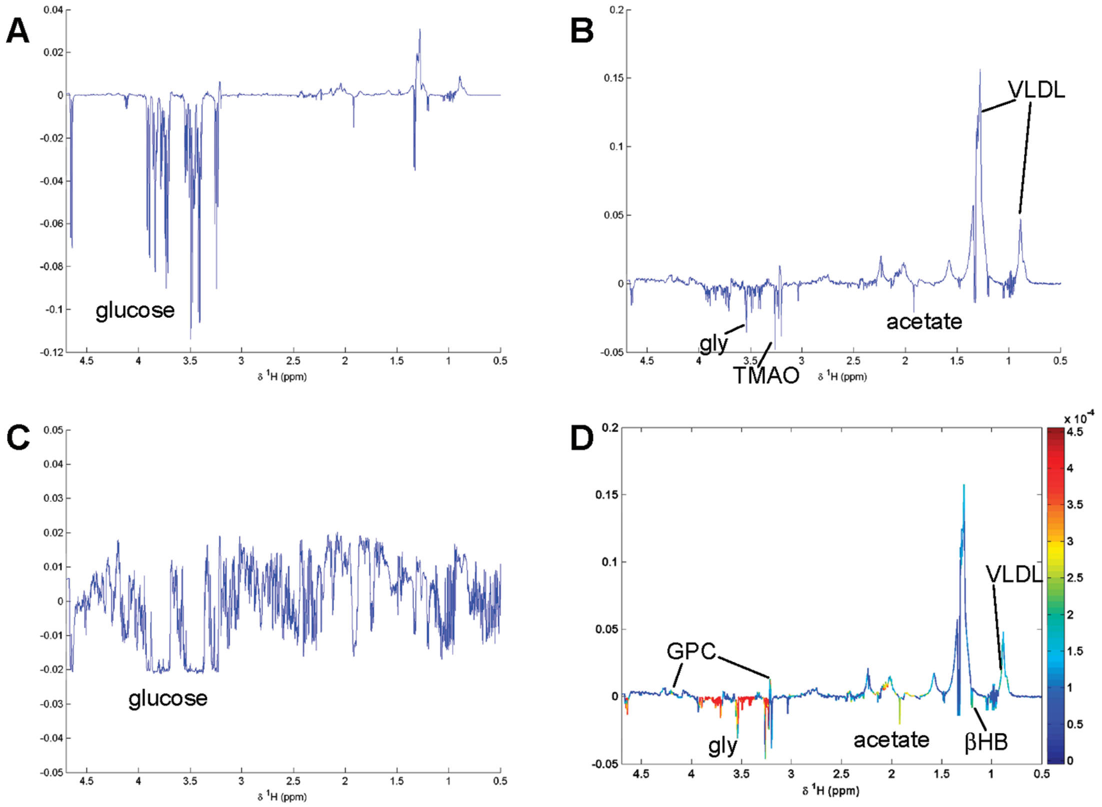


Figure 3. OPLS-DA of the high-dose group versus control group in the streptozotocin study. (A) Loadings from the OPLS-DA model constructed from the original data set. Note how the model is dominated by glucose. (B) Loadings of the OPLS-DA model constructed from the data set after the STOCSY^S procedure. gly = glycine, VLDL = very-low-density lipoprotein, TMAO = trimethylamine *N*-oxide, and GPC = glycerophosphocholine. (C) The loadings from the OPLS-DA model based on the original data set after UV scaling are difficult to interpret because of peak shape distortions. (D) Same model as in (C) after projection of the loadings as a color onto the covariance data. Again glucose dominates the loadings plots.

When glucose and lactate resonances are removed by simple excision of the variables from the data matrix, a similar result is observed (Figure 2 C), with samples from the high-dose group separating from the low-dose and controls groups mostly in PC2. However, the contributions of TMAO and choline could not be detected, because these resonances were so close to the glucose resonances at δ 3.24 that they had been excised from the data.

While PCA is useful for observing overall trends and identifying outliers, OPLS-DA can be used to find the quantitative relationship between the NMR data and the class of sample. The results from PCA suggested that animals in the high-dose group showed the largest degree of response to the toxin, with the exception of the nonresponder, an animal that did not appear hyperglycemic in response to 60 mg kg^{-1} streptozotocin. Therefore, two-class OPLS-DA models were constructed, between "high-dose" and "control" groups, after exclusion of the "low-dose" group. We compared the results from NMR data without scaling, with STOCSY^S, and with UV-scaling. A highly predictive model was constructed from the unscaled data, as judged by the high Q^2 (0.75). The robustness of this Q^2 was evaluated by 1000-fold response permutation, with $p \ll 1 \times 10^{-5}$ for the correlation between the original and permuted Q^2 (Figure S2A, Supporting Information). Figure 3 A shows the loadings plot for this model, indicating the variables that were most influential. This plot is clearly dominated by glucose resonances, which obscure the contributions from other metabolites. Figure 3 B shows the loadings plot for the OPLS-DA model from the STOCSY^S data. Again, a highly predictive model was constructed (Q^2 = 0.57, response permutation $p \ll 1 \times 10^{-5}$, Figure S2B). However, since the variables corresponding to glucose and lactate have been attenuated, the loadings plot is much more informative, and new biomarkers can be identified: elevated levels of TMAO, glycine, and acetate and decreased VLDL are associated with the high-dose group. The STOCSY^S procedure shows biomarker-like NMR peak shapes in the loadings plot, since almost all peaks in the spectrum retain their original Lorentzian line shape and coupling patterns, facilitating interpretation. For comparison, an OPLS-DA model was constructed from the same data but after UV-scaling, the scaling most commonly employed prior to OPLS-DA in metabonomics studies. Again, a highly predictive model was constructed (Q^2 = 0.65, $p \ll 1 \times 10^{-5}$, Figure S2C). The loadings plot in Figure 3 C indicates that the model was dominated by glucose, but interpretation is severely hampered by distortion of the shapes of the peaks and the amplification of the noise as a result of the UV-scaling procedure. A well-established method for increasing the interpretability of UV-scaled chemometrics models is to back-project the loadings as a color onto the covariance data.[35] In Figure 3D we have plotted the loadings from the UV-scaled model onto the loadings from the STOCSY^S model, effectively combining the results in parts B and C of Figure 3. This plot can help reveal biomarkers in the data set that were otherwise obscured by the glucose resonances, such as the decrease of glycerophosphocholine (GPC) in the high-dose group and the increase in glycine.

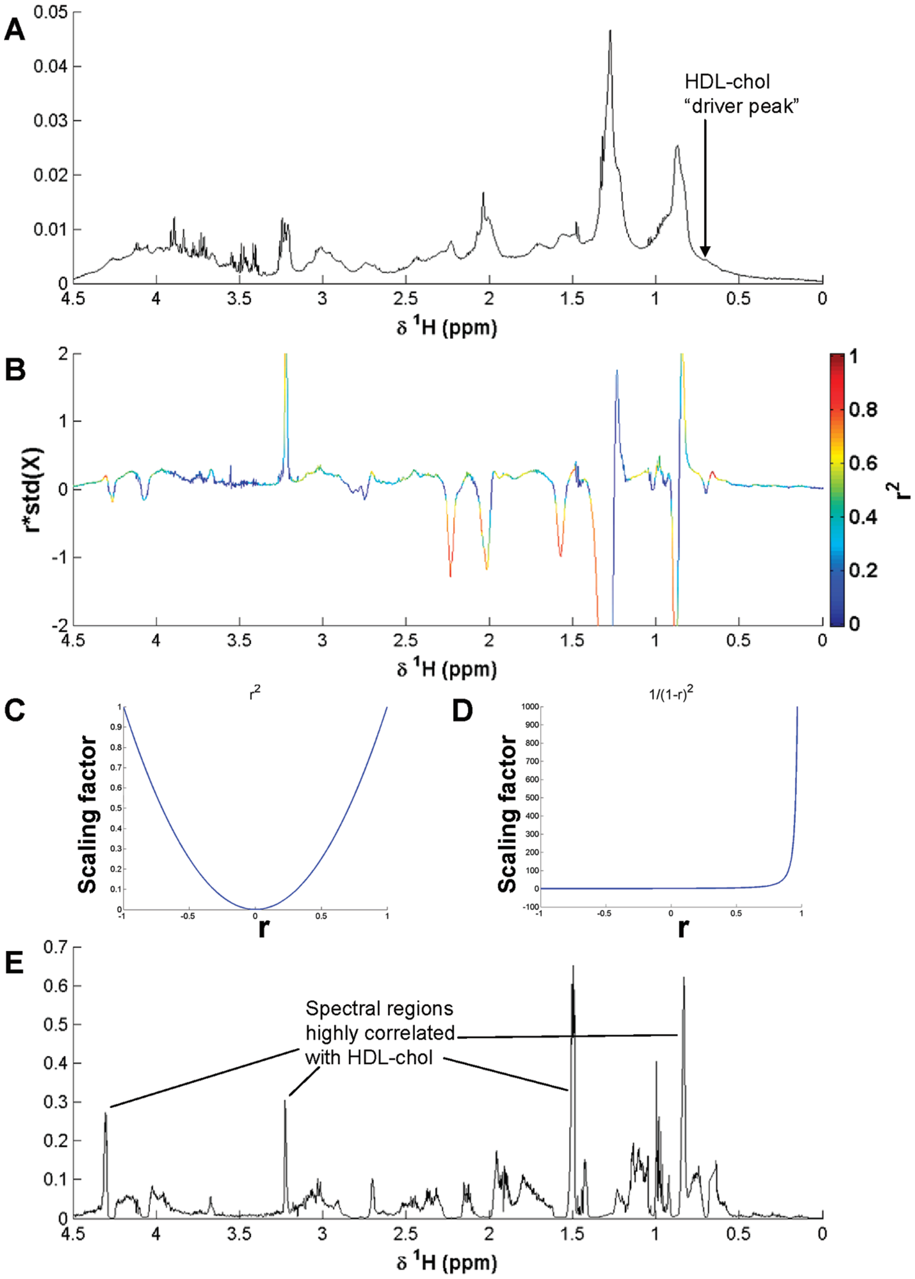


Figure 4. STOCSY^S enhancement of HDL-chol in plasma spectra from the MetS cohort study. (A) ^{1}H NMR spectrum from a typical human blood plasma sample from the MetS cohort. The arrow points to the HDL-chol resonance at δ 0.66 that was used as the driver peak for STOCSY. (B) The correlation coefficients of the NMR data with the peak at δ 0.66 are projected as a color onto the covariance data. (C) Graph of r^2 as a function of r. (C) Graph of $1/(1 - r^2)$ as a function of r. (E) The spectrum in (A) after STOCSY^S reveals enhancements of spectral regions highly correlated to the HDL-chol peak at δ 0.66.

Exemplification Using Human Plasma NMR Spectra. We have shown above that STOCSY^S can be used to remove resonances from metabolites that are not of interest. This was achieved by subtracting the variation correlated with glucose (and lactate) from the original data set, which can be denoted $\mathbf{X}_{\mathrm{full}}$. This can be considered as separating the data into correlated variables ($\mathbf{X}_{\mathrm{corr}}$) and uncorrelated variables ($\mathbf{X}_{\mathrm{uncorr}}$), viz.

$$\mathbf{X}_{\mathrm{full}} \rightarrow \mathbf{X}_{\mathrm{corr}} + \mathbf{X}_{\mathrm{uncorr}} \qquad (2)$$

In this equation, $\mathbf{X}_{\mathrm{uncorr}}$ is the scaled data used as input for the pattern recognition analysis as above. However, it is also possible to analyze $\mathbf{X}_{\mathrm{corr}}$ to investigate the variation in $\mathbf{X}_{\mathrm{full}}$ that is correlated with the main chosen peak of interest, the STOCSY driver peak. Each row in $\mathbf{X}_{\mathrm{corr}}$ resembles the pure compound spectrum from the selected metabolite or set of metabolites correlating with the selected data point. In this section we explore the potential for enhanced information recovery in metabonomic data sets that may be achieved by analyzing the information in $\mathbf{X}_{\mathrm{corr}}$.

Metabolic syndrome (MetS) is a constellation of risk factors that appear to promote cardiovascular disease and type 2 diabetes. Among these risk factors, atherogenic dyslipidemia has been postulated as being directly linked to development of atherosclerosis.[36] Although several organizations have offered slightly different criteria for diagnosis of MetS, all include

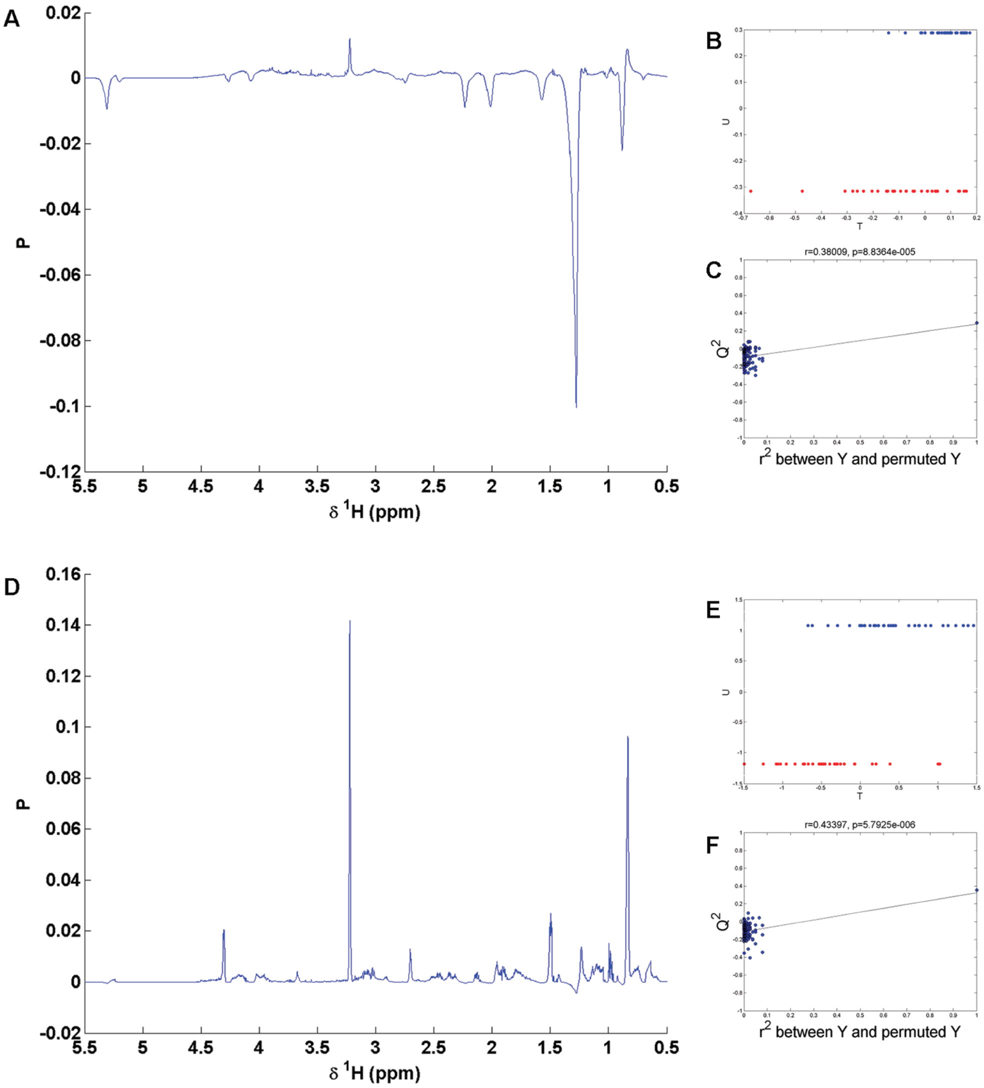


Figure 5. OPLS-DA models constructed from NMR spectra of plasma samples in the MetS cohort study. (A) Loadings plots from the OPLS-DA model of the unscaled 1D ^{1}H NMR data. (B) Inner relation plot for the model in (A) (T versus U). (C) Response permutation results for the OPLS-DA model in (A). (D) Loadings plot for the OPLS-DA model based on the data after STOCSY^S enhancement, driven from the HDL-chol peak at δ 0.66. (E) Inner relation plot (T versus U) for the model in (D). (F) Response permutation results for the model in (D).

elevated high-density lipoprotein cholesterol (HDL-chol). In the ^{1}H NMR spectrum of human blood plasma, a resonance at δ 0.66 has been assigned to "mostly" HDL-chol.[37] However, this peak is usually very low in intensity and consequently has only a minor influence in pattern recognition models. Therefore, the potential of STOCSY^S to enhance the weighting of this resonance and those correlated with it was explored.

Figure 4 A shows a typical ^{1}H NMR spectrum of a blood plasma sample from a patient with metabolic syndrome from the MetS data set. The arrow indicates the HDL-chol resonance at δ 0.66. Figure 4 B shows the covariance and correlation coefficients of the data points in the NMR spectra with the selected HDL-chol resonance peak in the form of a traditional STOCSY plot.[18] The plot reveals many spectral regions highly correlated to the resonance at δ 0.66, some positively and some negatively correlated. In the previous section, the square of the correlation coefficient (r^2) was used to attenuate signals from glucose and lactate. This relationship is plotted in Figure 4 C. To isolate and enhance the positively correlated variables, we have developed a new equation to transform the data set, namely

$$\mathbf{X}(\mathrm{sc})_{ij} = \mathbf{X}_{ij}/(1-\mathbf{r}_i)^2 \qquad (3)$$

which is plotted in Figure 4 D. The solution approaches 0.25 as r approaches -1 and approaches infinity as r approaches +1. Thus, spectral regions with correlations very close to 1 will be greatly enhanced, while negative correlations will be relatively attenuated (note it is necessary to remove correlations at exactly $r = 1$ to avoid singularities). Hence, spectral regions that are highly correlated with the HDL-chol peak at δ 0.66, such as the $N(CH_3)_3$ peak at δ 3.22, which is mainly from the choline-containing phospholipids present in HDL,[37] are highly emphasized in the transformed spectrum (Figure 4 E). We can hypothesize that the other greatly enhanced regions at δ 0.83, δ 1.50, and δ 4.30 are therefore also from HDL. These would otherwise be difficult to identify as they are confounded with other broad spectral features.

Figure 5 contrasts the OPLS-DA analyses of the metabolic syndrome data set with and without enhancement of HDL by the STOCSY^S procedure. Figure 5 A shows the loadings plot from OPLS-DA of the original 1D ^{1}H NMR data from 67 participants with and without MetS. Figure 5 B plots the inner relation (T versus U) for this OPLS-DA model, showing a high amount of overlap for the OPLS-DA scores. The Q^2 for this model was 0.29, and although this could be considered low, a 100-fold response permutation suggests that this predictive value was significant ($p = 8.8 \times 10^{-5}$). By contrast, parts D−F of Figure 5 show the results of the OPLS-DA model constructed from the STOCSY^S-transformed data. The inner relation plot now shows enhanced separation of the two groups (Figure 5 E), while Q^2 has increased to 0.35, with the response permutation $p = 5.8 \times 10^{-6}$ (Figure 5 F). Figure 5 D reveals that the most influential peaks were the $^+N(CH_3)_3$ choline headgroup resonances at δ 3.22 and other broad features from lipids at δ 0.83, δ 1.50, and δ 4.30. The identification of the resonance at δ 3.22 as most influential in identifying cases of MetS is consistent with a recent report.[38]

■ CONCLUSIONS

In this study we have shown that STOCSY^S can be used to suppress or enhance variation to modify the covariance patterns in the data that are modeled with multivariate statistics, enabling emphasis of more interesting parts of a data set. This can be considered a generalization of the STOCSY-editing approach introduced by Sands et al.[16] and can be readily applied to any data set in which known variation obscures more interesting features. However, an important advantage of STOCSY^S is that it can be used to extract biologically relevant features from spectral data without the need to define a cutoff value for any correlation coefficients.

The two presented examples illustrate the two main advantages of this technique. The first benefit is enhanced recovery of latent biochemical information. When large variation exists in a data set, this will dominate the principal component analysis and may obscure more interesting features. Removal of glucose peaks using STOCSY^S is an elegant alternative to simple excision of those spectral regions. This scaling is particularly useful if the dominating variation is from an unknown molecule, e.g., an ingested drug or its metabolite. The scaling enhancement can, for example, be useful to study an endogenous molecule from a patient with an unknown inborn error of metabolism, where the chemical signature of spectral regions containing the molecule is unknown. The second advantage is the improved interpretability of pattern recognition models. By enhancing spectral regions correlated with a known peak of interest, the data supplied for pattern recognition are more likely to have discernible levels of discriminating molecules, which benefits sample classification and model interpretation.

In our first example data set, we showed the usefulness of STOCSY^S in a toxicological study on streptozotocin administration. Although we effectively suppressed signals from glucose and lactate, STOCSY^S will also suppress other metabolites that are correlated with these signals. For example, a comparison of the biomarkers identified in the PC2 loadings in Figure 2 B with Figure 2 C shows that the influence of acetate in the STOCSY^S data was reduced compared to the data in which variables corresponding to glucose and lactate were removed by the eye. This is because the acetate peak intensity is correlated with that of the lactate peak in this data set (r = 0.8). Importantly, even the "suppressed" peak intensities will still be directly proportional to the molecular concentration in the sample. Thus, it still contributes to the separation between the groups and can be readily identified in the loadings plot. It should be noted that some metabolites will covary across samples in an NMR data set, and as a result, applying STOCSY^S to one metabolite would necessarily attenuate the signals of the other. There are, of course, many situations in which this is the desired result, such as attenuation of all amino acid resonances in a time-course data set from a seminal fluid sample.[39] In this situation, signals from many amino acids being cleaved off seminogellin proteins are covarying and could potentially be attenuated after one round of STOCSY^S. However, if the aim is to suppress the signal from one specific amino acid, it should be possible to add extra variation into the model by concatenating several time courses, which show different covariation patterns and will therefore allow discrimination of signals from different amino acids.

A potential limitation to the success of STOCSY and all of its derivative methods is the presence of peak positional variation. Although in mammalian plasma samples the homeostatic control over osmolarity and pH minimizes this problem, it is still prevalent in urinary data sets. This can be ameliorated to some extent with the development of increasingly sophisticated alignment methods.[9,40] Alternatively, dimensionality reduction techniques such as targeted profiling[41] or deconvolution may be explored.[42] Equation 3 for enhancement of spectral features is beneficial because it approaches an asymptote at $r = +1$ (as opposed to, e.g., multiplication of $\mathbf{X}_{ij}$ by $\mathbf{r}_i^2$) and thus greatly emphasizes spectral features highly correlated to the peak of interest. However, it should be noted that several other similar equations could be used in this place, and variants of this equation will be explored in future work.

The STOCSY^S approach is easy to implement and is valuable for diminishing the influence of unwanted spectral variation and enhancing regions of interest. This approach can be applied to any quantitative data set to prevent the dominance of high-intensity variables and emphasize low-intensity variables and can find wide application within the metabolic NMR community as well as in bioanalytical NMR research and other studies on complex mixtures.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional information as noted in text. This material is available free of charge via the Internet at http://pubs.acs.org

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: a.maher@imperial.ac.uk (A.D.M.); j.nicholson@imperial.ac.uk (J.K.N.).

■ ACKNOWLEDGMENTS

A.D.M. acknowledges funding from the EU FP6MolPAGE project (Grant LSHG-512066). This work was supported by the award of an RSC Ph.D. studentship to J.M.F. and by the Australian National Health and Medical Research Council (Project Grant 568767) to C.D.R. We are grateful to the COMET project group for funding and providing data.[43,44] The MRC Integrative Toxicology Training Partnership (ITTP) is acknowledged for financial support to M.C.

■ REFERENCES

(1) Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E. *Nat. Rev. Drug Discovery* **2002**, *1*, 153−161.

(2) Wilson, I. D.; Plumb, R.; Granger, J.; Major, H.; Williams, R.; Lenz, E. A. *J. Chromatogr., B* **2005**, *817*, 67−76.

(3) Trygg, J.; Holmes, E.; Lundstedt, T. *J. Proteome Res.* **2007**, *6*, 469−479.
(4) Lloyd, G. R.; Wongravee, K.; Silwood, C. J. L.; Grootveld, M.; Brereton, R. G. *Chemom. Intell. Lab. Syst.* **2009**, *98*, 149−161.
(5) Fonville, J. M.; Richards, S. E.; Barton, R. H.; Boulange, C. L.; Ebbels, T. M. D.; Nicholson, J. K.; Holmes, E.; Dumas, M. E. *J. Chemom.* **2011**, *24*, 636−649.
(6) Carr, H. Y.; Purcell, E. M. *Phys. Rev.* **1954**, *94*, 630.
(7) Fonville, J. M.; Maher, A. D.; Coen, M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2010**, *82*, 1811−1821.
(8) Maher, A. D.; Crockford, D.; Toft, H.; Malmodin, D.; Faber, J. H.; McCarthy, M. I.; Barrett, A.; Allen, M.; Walker, M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2008**, *80*, 7354−7362.
(9) Veselkov, K. A.; Lindon, J. C.; Ebbels, T. M.; Crockford, D.; Volynkin, V. V.; Holmes, E.; Davies, D. B.; Nicholson, J. K. *Anal. Chem.* **2009**, *81*, 56−66.
(10) Craig, A.; Cloarec, O.; Holmes, E.; Nicholson, J. K.; Lindon, J. C. *Anal. Chem.* **2006**, *78*, 2262−2267.
(11) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis: Principles and Applications*; Umetrics: Umea, Sweden, 2001.
(12) Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Prog. NMR Spectrosc.* **2001**, *39*, 1−40.
(13) Trygg, J.; Wold, S. *J. Chemom.* **2002**, *16*, 119−128.
(14) Wold, S.; Antti, H.; Lindgren, F.; Ohman, J. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 175−185.
(15) Kemsley, E. K.; Tapp, H. S. *J. Chemom.* **2009**, *23*, 263−264.
(16) Sands, C. J.; Coen, M.; Maher, A. D.; Ebbels, T. M. D.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2009**, *81*, 6458−6466.
(17) Bales, J. R.; Higham, D. P.; Howe, I.; Nicholson, J. K.; Sadler, P. J. *Clin. Chem.* **1984**, *30*, 426−432.
(18) 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*, 1282−1289.
(19) Noda, I.; Dowrey, A. E.; Marcott, C.; Story, G. M.; Ozaki, Y. *Appl. Spectrosc.* **2000**, *54*, 236A−248A.
(20) Bruschweiler, R.; Zhang, F. L. *J. Chem. Phys.* **2004**, *120*, 5253−5260.
(21) Holmes, E.; Loo, R. L.; Cloarec, O.; Coen, M.; Tang, H.; Maibaum, E.; Bruce, S.; Chan, Q.; Elliott, P.; Stamler, J.; Wilson, I. D.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2007**, *79*, 2629−2640.
(22) Coen, M.; Hong, Y. S.; Cloarec, O.; Rhode, C. M.; Reily, M. D.; Robertson, D. G.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2007**, *79*, 8956−8966.
(23) Keun, H. C.; Athersuch, T. J.; Beckonert, O.; Wang, Y.; Saric, J.; Shockcor, J. P.; Lindon, J. C.; Wilson, I. D.; Holmes, E.; Nicholson, J. K. *Anal. Chem.* **2008**, *80*, 1073−1079.
(24) Crockford, D. J.; Maher, A. D.; Ahmadi, K. R.; Barrett, A.; Plumb, R. S.; Wilson, I. D.; Nicholson, J. K. *Anal. Chem.* **2008**, *80*, 6835−6844.
(25) Blaise, B. J.; Navratil, V.; Emsley, L.; Toulhoat, P. *J. Proteome Res.* **2011**, *10*, 4342−4348.
(26) Guyett, P.; Glushka, J.; Gu, X. G.; Bar-Peled, M. *Carbohydr. Res.* **2009**, *344*, 1072−1078.
(27) Kirwan, G. M.; Coffey, V. G.; Niere, J. O.; Hawley, J. A.; Adams, M. J. *Anal. Chim. Acta* **2009**, *652*, 173−179.
(28) Sands, C. J.; Coen, M.; Ebbels, T. M. D.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2011**, *83*, 2075−2082.
(29) Lindon, J. C.; Keun, H. C.; Ebbels, T. M.; Pearce, J. M.; Holmes, E.; Nicholson, J. K. *Pharmacogenomics* **2005**, *6*, 691−699.
(30) Meiboom, S.; Gill, D. *Rev. Sci. Instrum.* **1958**, *29*, 688−691.
(31) Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(32) Zhang, S.; Nagana Gowda, G. A.; Asiago, V.; Shanaiah, N.; Barbas, C.; Raftery, D. *Anal. Biochem.* **2008**, *383*, 76−84.
(33) van Doorn, M.; Kemme, M.; Ouwens, M.; van Hoogdalem, E. J.; Jones, R.; Romijn, H.; de Kam, M.; Schoemaker, R.; Burggraaf, K.; Cohen, A. *Br. J. Clin. Pharmacol.* **2006**, *62*, 391−402.
(34) Couto Alves, A.; Rantalainen, M.; Holmes, E.; Nicholson, J. K.; Ebbels, T. M. D. *Anal. Chem.* **2009**, *81*, 2075−2084.
(35) Cloarec, O.; Dumas, M. E.; Trygg, J.; Craig, A.; Barton, R. H.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *Anal. Chem.* **2005**, *77*, 517−526.
(36) Grundy, S. M.; Cleeman, J. I.; Daniels, S. R.; Donato, K. A.; Eckel, R. H.; Franklin, B. A.; Gordon, D. J.; Krauss, R. M.; Savage, P. J.; Smith, S. C. Jr.; Spertus, J. A.; Costa, F. *Circulation* **2005**, *112*, 2735−2752.
(37) Nicholson, J. K.; Foxall, P. J.; Spraul, M.; Farrant, R. D.; Lindon, J. C. *Anal. Chem.* **1995**, *67*, 793−811.
(38) Makinen, V. P.; Soininen, P.; Forsblom, C.; Parkkonen, M.; Ingman, P.; Kaski, K.; Groop, P. H.; Ala-Korpela, M. *Mol. Syst. Biol.* **2008**, *4*, 167.
(39) Maher, A. D.; Cloarec, O.; Patki, P.; Craggs, M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2009**, *81*, 288−295.
(40) Alm, E.; Torgrip, R. J. O.; Aberg, K. M.; Schuppe-Koistinen, I.; Lindberg, J. *Anal. Bioanal. Chem.* **2009**, *395*, 213−223.
(41) Weljie, A. M.; Newton, J.; Mercier, P.; Carlson, E.; Slupsky, C. M. *Anal. Chem.* **2006**, *78*, 4430−4442.
(42) Chylla, R. A.; Hu, K.; Ellinger, J. J.; Markley, J. L. *Anal. Chem.* **2011**, *83*, 4871−4880.
(43) Lindon, J. C.; Keun, H. C.; Ebbels, T. M. D.; Pearce, J. M. T.; Holmes, E.; Nicholson, J. K. *Pharmacogenomics* **2005**, *6*, 691−699.
(44) Ebbels, T. M. D.; Keun, H. C.; Beckonert, O. P.; Bollard, M. E.; Lindon, J. C.; Holmes, E.; Nicholson, J. K. *J. Proteome Res.* **2007**, *6*, 4407−4422.