Journal of Magnetic Resonance **154,** 163–175 (2002)
doi:10.1006/jmre.2001.2486, available online at http://www.idealibrary.com on IDEAL®

NMR Spectral Quantitation by Principal Component Analysis

III. A Generalized Procedure for Determination of Lineshape Variations

R. Stoyanova* and T. R. Brown†,[1]

*Division of Population Science, Fox Chase Cancer Center, 7701 Burholme Avenue, Philadelphia, Pennsylvania 19111; and †Hatch Center for MR Research, Departments of Radiology and Biomedical Engineering, Columbia University, 710 West 168th Street, New York, New York 10032

Received April 7, 2000; revised November 6, 2001

We present a general procedure for automatic quantitation of a series of spectral peaks based on principal component analysis (PCA). PCA has been previously used for spectral quantitation of a single resonant peak of constant shape but variable amplitude. Here we extend this procedure to estimate all of the peak parameters: amplitude, position (frequency), phase and linewidth. The procedure consists of a series of iterative steps in which the estimates of position and phase from one stage of iteration are used to correct the spectra prior to the next stage. The process is convergent to a stable result, typically in less than 5 iterations. If desired, remaining linewidth variations can then be corrected. Correction of (typically) unwanted variations of these types is important not only for direct peak quantitation, but also as a preprocessing step for spectral data prior to application of pattern recognition/classification techniques. The procedure is demonstrated on simulated data and on a set of 992 ^{31}P NMR *in vivo* spectra taken from a kinetic study of rat muscle energetics. The proposed procedure is robust, makes very limited assumptions about the lineshape, and performs well with data of low signal-to-noise ratio. © 2002 Elsevier Science (USA)

Key Words: **spectroscopy; principal component analysis; automatic quantitation; peak parameters; lineshape correction.**

INTRODUCTION

Sophisticated and accurate approaches to quantitation of resonance peaks in a single spectrum (in either the time or frequency domain) are widely available in the NMR community. However, in recent years new instrumental procedures able to acquire hundreds of related spectra in a short time have generated a need for approaches which can analyze an entire set of such related spectra as a whole, thus taking advantage of the relationships among the spectra to improve the quality of the analysis. Approaches that do this are particularly useful for spectra with low signal-to-noise ratio (SNR) as they utilize the collective power of the data.

Principal component analysis (PCA) is one such general statistical method, which successfully estimates the amplitude of a constant resonance peak across a series of spectra (*1–7*). Typically, in real data there are numerous peak imperfections, caused by field inhomogeniety, susceptibility effects, and other instrumental/experimental conditions. Variations in the peak lineshape, such as frequency, phase, or linewidth, can contribute significant error to the area estimations (*2*). In the previous paper in this series (*2*) we presented an iterative PCA-based automatic method for determining frequency and phase variations among the spectra in large spectral datasets and if necessary, subsequently correcting the data for such variations. The approach equated the PCA-based data decomposition with the first-order Taylor expressions of the peak shape with respect to frequency and phase variations and then solved these vector equations by projecting them onto the subspace defined by the first three principal components. In general, this approach was necessarily iterative because the Taylor series expansion was only first order and required only small variations to be able to estimate them accurately. Thus after the spectra were corrected for their estimated frequency and phase variations, a next round of PCA decomposition was carried out. This approach lacked clear stopping criteria since the second- and higher order principal components (PCs) vanished as the variations were removed by the successive iterations. This was noted by us (*7*) and others (*5*). Recently, Witjes *et al.* (*5*), also on the basis of a Taylor series expansion, proposed a linear least-squares approach to estimating the phase and frequency variations by projecting the individual spectra onto functions derived from the estimated lineshape as determined from the shape of the first PC. This resulted in an iterative procedure which converged to the correct answer for frequency variations of the order of the linewidth but a limited range of phase variations ($\pm 60^\circ$). A novel approach to linewidth correction by modeling broad lines as the sum of frequency shifted versions of the first PC, following the removal of frequency and phase variations, was also presented in (*5*).

Here we present a generalized approach to simultaneous quantification of all peak characteristics, amplitude, frequency,

[1] To whom correspondence should be addressed. Fax: (212) 305-0175. E-mail: trb11@columbia.edu.

1090-7807/02 $35.00

phase, and linewidth, using an improved PCA decomposition of each spectrum. The procedure includes novel modeling of the linewidth variations and is superior to those reported previously in terms of the range of variations it can determine, its stability, and convergence. The procedure is demonstrated on simulated data and on a set of 992 ^{31}P NMR *in vivo* spectra taken from a kinetic study of rat muscle energetics.

THEORY

PCA is a well-known statistical technique and has been used extensively to analyze large multidimensional datasets (*8*). It identifies fundamental structures in a data matrix, called principal components, through an orthogonal decomposition of the data, using the PCs ($\vec{\mathbf{P}}_1, \vec{\mathbf{P}}_2, \ldots$) as a basis set,

$$\mathbf{D} = S_1\vec{\mathbf{P}}_1 + S_2\vec{\mathbf{P}}_2 + S_3\vec{\mathbf{P}}_3 + \cdots + S_m\vec{\mathbf{P}}_m, \qquad [1]$$

where $\mathbf{D}$ is the data matrix ($n \times m$, $m \leq n$), $\vec{\mathbf{P}}_j$ are ($1 \times m$) matrices that can be thought as m-dimensional vectors, and S_j are ($n \times 1$) matrices, which are the projections of the data along the PCs, generally called scores ($j = 1, \ldots, m$). Thus we represent $\mathbf{D}$ as the sum of products of ($n \times 1$) times ($1 \times m$) matrices. Equation [1] can also be written in matrix form as

$$\mathbf{D} = \mathbf{SP}, \qquad [2]$$

where the rows of $\mathbf{P}$ are the $\vec{\mathbf{P}}_j$ and the columns of $\mathbf{S}$ are the projections of $\mathbf{D}$ on $\vec{\mathbf{P}}_j$. To calculate $\mathbf{S}$ and $\mathbf{P}$ we need to diagonalize the covariance matrix $\mathbf{D^T D}$ (we assume $\mathbf{D}$ is real):

$$\frac{1}{m}\mathbf{D^T DQ} = \mathbf{Q}\Lambda. \qquad [3]$$

$\mathbf{Q}$ is the matrix of the eigenvectors of $\mathbf{D^T D}$ and Λ is a diagonal matrix of its eigenvalues, ordered by amplitude. Since $\mathbf{D^T D}$ is symmetric, the eigenvalues are real and nonnegative. The PCs, $\vec{\mathbf{P}}_1, \vec{\mathbf{P}}_2, \ldots$, are the transposed eigenvectors; i.e., $\mathbf{P} = \mathbf{Q}^{\mathrm{T}}$. Since $\mathbf{D} = \mathbf{SP}$ and $\mathbf{PP^T} = \mathbf{I}$, the scores $\mathbf{S} = \mathbf{DP^T}$.

Let $\mathbf{D}$ be a spectral dataset of n spectra, m points each, in which the only coherent variation is the amplitude of a fixed single lineshape $\bar{\mathbf{f}}(\omega_j)$; i.e., each spectrum consists of the line $\bar{\mathbf{f}}(\omega_j)$ of arbitrary magnitude plus random noise (assumed to have the same variance for all spectra). In (*1*) we showed that the normalized lineshape $\bar{\mathbf{f}}$ of the peak is well approximated by the first PC,

$$\bar{\mathbf{f}} = \frac{\vec{\mathbf{P}}_1}{\sum_{j=1}^{m} P_{1j}} \qquad [4]$$

while the amplitude of the peak in each spectrum can be estimated by

$$A = \left(\sum_{j=1}^{m} P_{1j}\right) S_1, \qquad [5]$$

where $A = (A_1, A_2, \ldots, A_n)$ are the amplitudes of $\bar{\mathbf{f}}$ in each spectrum in $\mathbf{D}$ and P_{1j} is the component of $\vec{\mathbf{P}}_1$ at frequency ω_j. In fact, the estimates A calculated in this way are the best global, linear estimates for the true amplitudes in the sense that their mean square error over the entire dataset is minimized (*1*).

In general, however, variations in peak shape and position exist in real data. Further, since an NMR spectrum is acquired by sine and cosine projection from a complex waveform, variation in phase between spectra must also be considered. Thus a full analysis must be able to take into account additional variations beyond the amplitude. In (*2*) we showed that variations in the phase and position of the peak resulted in a decomposition of $\mathbf{D}$ in Eq. [1] in which the presence of phase and frequency variations of the peaks was reflected in the shapes of the PCs while the scores estimated their strengths in each spectrum. We made no attempt in (*2*) to address changes in peak shape caused by linewidth variations. In general, however, variations in the linewidth of the peak across series of spectra are also present. Thus to address the general case we need to consider a lineshape $\bar{\mathbf{f}}(\omega_j)$ which may have varying amplitude, frequency, phase, and linewidth throughout the dataset. Amplitude, frequency, and phase variations are well defined for any peak shape since they involve fixed transformations of an arbitrary peak shape. Amplitude corresponds to the strength of the shape in each spectrum, frequency to the shift of the shape along the frequency axis, and phase to the projection of the complex waveform along a different direction in the complex plane. Variations in linewidth, however, present a more difficult problem since we need to characterize how the peak shape changes with its variation. Thus, we assume for our further analysis that the peak shape is some well parameterized function of its linewidth (lorenzian, gaussian, etc.).

Under these conditions we then have a series of spectra, $A_i\bar{\mathbf{f}}(\omega_j - \omega_i, \tau_i, \varphi_i)$, where for the ith spectrum, A_i is the peak amplitude; $\omega_i = \omega_0 + \delta\omega_i$ is the frequency offset parametized by $\delta\omega_i$, the shift from the average frequency position ω_0; $\tau_i = \tau_0 + \delta\tau_i$ is the linewidth parameter with $\delta\tau_i$ fluctuation from the average linewidth τ_0; and φ_i is the phase (relative to the entire dataset). Let $A, \delta\omega, \delta\tau$, and φ be ($n \times 1$) matrices containing $A_i, \delta\omega_i, \delta\tau_i$, and φ_i ($i = 1, \ldots, n$). To determine them, we expand $\bar{\mathbf{f}}$ in Taylor series in the frequency and linewidth variations since they are generally small (*7*). Also we used the imaginary part of the signal $\bar{\mathbf{f}}^{\mathrm{I}}$ to model the variations in phase (here and in the rest of the paper we will use $\bar{\mathbf{f}}^{\mathrm{I}}$ to denote the imaginary (or dispersive part) of the signal and $\bar{\mathbf{f}}$ will refer only to the real (or absorptive part)).

We consider only the first-order terms in the Taylor series since we assume $\delta\omega$, $\delta\tau$, and φ are small. Thus we obtain

$$S_1\vec{\mathbf{P}}_1 + S_2\vec{\mathbf{P}}_2 + S_3\vec{\mathbf{P}}_3 + S_4\vec{\mathbf{P}}_4 = A\left[\vec{\mathbf{f}}\cos\varphi + \frac{\partial\vec{\mathbf{f}}}{\partial\omega}\bigg|_{\omega_0}\delta\omega\cos\varphi + \frac{\partial\vec{\mathbf{f}}}{\partial\tau}\bigg|_{\tau_0}\delta\tau\cos\varphi - \vec{\mathbf{f}}^{\mathrm{I}}\sin\varphi\right] + \bar{\varepsilon} \quad [6]$$

or

$$[\vec{\mathbf{P}}_1\ \vec{\mathbf{P}}_2\ \vec{\mathbf{P}}_3\ \vec{\mathbf{P}}_4]\begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix} = \begin{bmatrix} \vec{\mathbf{f}} & \frac{\partial\vec{\mathbf{f}}}{\partial\omega} & \frac{\partial\vec{\mathbf{f}}}{\partial\tau} & \vec{\mathbf{f}}^{\mathrm{I}} \end{bmatrix}\begin{bmatrix} A\cos\varphi \\ A\,\delta\omega\cos\varphi \\ A\,\delta\tau\cos\varphi \\ -A\sin\varphi \end{bmatrix} + \bar{\varepsilon}, \quad [7]$$

where $\bar{\varepsilon}$ represents all higher order terms and noise in the data. PCA of such a dataset will yield second- and higher order PCs that are mixtures of the shapes associated with the individual variations. For each spectrum this is a vector equation in the dimensional space spanned by the PCs. We solve this vector equation by projecting both sides onto the four vectors $\vec{\mathbf{f}}$, $\frac{\partial\vec{\mathbf{f}}}{\partial\omega}$, $\frac{\partial\vec{\mathbf{f}}}{\partial\tau}$, and $\vec{\mathbf{f}}^{\mathrm{I}}$, which we construct using $\vec{\mathbf{P}}_1$. This yields a set of four matrix equations, which can be solved for the unknowns A, $\delta\omega$, $\delta\tau$, and φ,

$$\begin{bmatrix} \vec{\mathbf{P}}_1\cdot\vec{\mathbf{f}} & \vec{\mathbf{P}}_2\cdot\vec{\mathbf{f}} & \vec{\mathbf{P}}_3\cdot\vec{\mathbf{f}} & \vec{\mathbf{P}}_4\cdot\vec{\mathbf{f}} \\ \vec{\mathbf{P}}_1\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\omega} & \vec{\mathbf{P}}_2\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\omega} & \vec{\mathbf{P}}_3\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\omega} & \vec{\mathbf{P}}_4\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\omega} \\ \vec{\mathbf{P}}_1\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\tau} & \vec{\mathbf{P}}_2\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\tau} & \vec{\mathbf{P}}_3\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\tau} & \vec{\mathbf{P}}_4\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\tau} \\ \vec{\mathbf{P}}_1\cdot\vec{\mathbf{f}}^{\mathrm{I}} & \vec{\mathbf{P}}_2\cdot\vec{\mathbf{f}}^{\mathrm{I}} & \vec{\mathbf{P}}_3\cdot\vec{\mathbf{f}}^{\mathrm{I}} & \vec{\mathbf{P}}_4\cdot\vec{\mathbf{f}}^{\mathrm{I}} \end{bmatrix}\begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \end{bmatrix}$$

$$= \begin{bmatrix} \vec{\mathbf{f}}\cdot\vec{\mathbf{f}} & \frac{\partial\vec{\mathbf{f}}}{\partial\omega}\cdot\vec{\mathbf{f}} & \frac{\partial\vec{\mathbf{f}}}{\partial\tau}\cdot\vec{\mathbf{f}} & \vec{\mathbf{f}}^{\mathrm{I}}\cdot\vec{\mathbf{f}} \\ \vec{\mathbf{f}}\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\omega} & \frac{\partial\vec{\mathbf{f}}}{\partial\omega}\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\omega} & \frac{\partial\vec{\mathbf{f}}}{\partial\tau}\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\omega} & \vec{\mathbf{f}}^{\mathrm{I}}\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\omega} \\ \vec{\mathbf{f}}\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\tau} & \frac{\partial\vec{\mathbf{f}}}{\partial\omega}\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\tau} & \frac{\partial\vec{\mathbf{f}}}{\partial\tau}\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\tau} & \vec{\mathbf{f}}^{\mathrm{I}}\cdot\frac{\partial\vec{\mathbf{f}}}{\partial\tau} \\ \vec{\mathbf{f}}\cdot\vec{\mathbf{f}}^{\mathrm{I}} & \frac{\partial\vec{\mathbf{f}}}{\partial\omega}\cdot\vec{\mathbf{f}}^{\mathrm{I}} & \frac{\partial\vec{\mathbf{f}}}{\partial\tau}\cdot\vec{\mathbf{f}}^{\mathrm{I}} & \vec{\mathbf{f}}^{\mathrm{I}}\cdot\vec{\mathbf{f}}^{\mathrm{I}} \end{bmatrix}\begin{bmatrix} A\cos\varphi \\ A\delta\omega\ \cos\varphi \\ A\,\delta\tau\cos\varphi \\ -A\sin\varphi \end{bmatrix} \quad [8]$$

or

$$\begin{bmatrix} A\cos\varphi \\ A\,\delta\omega\cos\varphi \\ A\,\delta\tau\cos\varphi \\ -A\sin\varphi \end{bmatrix} = [\mathbf{F}^{\mathrm{T}}\cdot\mathbf{F}]^{-1}[\mathbf{F}^{\mathrm{T}}\cdot\mathbf{P}][S], \quad [9]$$

where $\mathbf{F}$ consists of $\vec{\mathbf{f}}$, $\frac{\partial\vec{\mathbf{f}}}{\partial\omega}$, $\frac{\partial\vec{\mathbf{f}}}{\partial\tau}$, and $\vec{\mathbf{f}}^{\mathrm{I}}$, and $\mathbf{P}$ and $\mathbf{S}$ contain the first four PCs and their scores, respectively. Note that we have assumed that the projections of $\mathbf{F}$ on $\bar{\varepsilon}$ are zero to solve this set of equations. This is true only if the variations are small.

Our previous paper (*2*) attempted to solve these equations (actually a subset which was missing the linewidth variations) by projections onto the PCs rather than vectors derived from $\vec{\mathbf{f}}$. As the variations were removed the higher PCs were increasingly dominated by noise and the resultant matrices became poorly defined. By projecting onto the vectors derived from $\vec{\mathbf{f}}$ the procedure is now well behaved.

Let S^{F} be a matrix, containing the transformed scores S from the right-hand side of Eq. [9]. Then the estimation of A, $\delta\omega$, $\delta\tau$, and φ is straightforward:

$$\begin{aligned} A &= \frac{S_1^{\mathrm{F}}}{\cos\left(-\arctan\frac{S_4^{\mathrm{F}}}{S_1^{\mathrm{F}}}\right)} \\ \delta\omega &= \frac{S_2^{\mathrm{F}}}{S_1^{\mathrm{F}}} \\ \delta\tau &= \frac{S_3^{\mathrm{F}}}{S_1^{\mathrm{F}}} \\ \varphi &= -\arctan\frac{S_4^{\mathrm{F}}}{S_1^{\mathrm{F}}} \end{aligned} \quad [10]$$

Equation [10] provides estimates of all peak parameters needed for a complete quantitation. In addition the estimates for frequency, linewidth, and phase variations can be used to correct the data.

METHODS

Since the proposed approach is applied to the real spectral signal in the frequency domain, it is asumed that the data are Fourier transformed prior to PCA. Other preprocessing of a routine time-to-frequency nature may be performed, such as phasing the entire dataset, applying a lorentzian filter, and zero-filling if desired. In the derivation of Eq. [10] we assumed small frequency and linewidth variations. In cases where the frequency shifts in the dataset are larger than the average peak linewidth, a crude alignment (for instance, by aligning the highest point in the peak) could be advantageous. The interpretation of the PC's shapes will be easier if the spectra are initially phased so that the real part resembles an absorption shape (through automatic procedures or by applying phase parameters, estimated from the sum of the spectra). Once the preprocessing is finished and the peak of interest isolated, PCA is applied directly to the intensities of the points from the spectral region containing the peak. This is computationally advantageous and also reduces potential interference from the neighboring peaks.

A detailed description of the PCA correction procedure is provided in the Appendix. To estimate the vectors in $\mathbf{F}$ we use Eq. 4 to estimate $\vec{\mathbf{f}}$. Given $\vec{\mathbf{f}}$ we calculate $\frac{\partial\vec{\mathbf{f}}}{\partial\omega}$ and $\vec{\mathbf{f}}^{\mathrm{I}}$ by numerical differentiation and Hilbert transform. In our particular implementation we perform a Hilbert transform by applying inverse FFT to $\vec{\mathbf{f}}$, swapping the real and imaginary parts and forward

FFT (alternatively, the Hilbert transform option can be used if available in the software package used). The calculation of $\frac{\partial \vec{\mathbf{f}}}{\partial \tau}$ is slightly more complex. We transform $\vec{\mathbf{f}}$ into the time domain, apply a lorentzian filter of 1 Hz, subtract the result from the original FID, and finally transform back to frequency space. Although exact only for lorentzian lines, this has proven to be sufficiently accurate when applied to real data. This is the step in the procedure which requires relatively smoothly varying lineshapes. Abrupt discontinuities would not be modeled well by this procedure. Fortunately lineshapes in NMR spectra are smoothly varying. In our implementation calculating both $\vec{\mathbf{f}}^{\mathrm{I}}$ and $\frac{\partial \vec{\mathbf{f}}}{\partial \tau}$ requires Fourier transformation of $\vec{\mathbf{f}}$ back to the time domain. Since $\vec{\mathbf{P}}_1$ is real, its Fourier transform to the time domain yields a symmetric FID. We zero half of the points and perform the required steps to calculate $\vec{\mathbf{f}}^{\mathrm{I}}$ and $\frac{\partial \vec{\mathbf{f}}}{\partial \tau}$ as described above. Because half of the points are zeroed, the final result needs to be scaled by a factor of 2. When an FFT is used for the above steps, care should be taken that $\vec{\mathbf{P}}_1$ is zero-filled to the closest power of 2.

Once the vectors in $\mathbf{F}$ are estimated, calculating the right-hand side of Eq. [9] is straightforward. The estimates in Eq. [10] provide a full peak quantitation for each spectrum in the dataset. Note that when the phase variations approach their maximum ($\pm 90^\circ$), small errors in the phase estimate can cause erroneous change of the sign of the estimated phase. Care must be taken to calculate the arc tangent properly; i.e., from $-180^\circ + 180^\circ$ rather than just -90° to $+90^\circ$. The estimates from Eq. [10] are relative to the average value of the parameters. For most practical applications measuring the relative changes in the peaks across the dataset is sufficient. However, the absolute values of the peak parameters can also be estimated. For example, the kth peak in the dataset has amplitude A_k, peak position $\omega_0 + \delta\omega_k$, linewidth $\tau_0 + \delta\tau_k$, and phase φ_k. ω_0 and τ_0 can be measured from $\vec{\mathbf{P}}_1$ manually.

The above estimates are then used to correct the spectra. For the kth spectrum the phase is adjusted by multiplying the complex data in frequency domain by $\exp(i\varphi_k)$; both corrections for frequency and linewidth variations are performed in time domain, where the lth point of the FID is multiplied by $\exp(i 2\pi\, \delta\omega_k\, l \Delta t)$ and $\exp(\pi\, \delta\tau_k\, l \Delta t)$, resp. ($\Delta t$ is the dwell time). Note that these correction factors are 1 for the first point of the FID and thus they do not change the peak amplitude. Since we ignored the higher order terms in Eq. [7], the correction procedure needs to be iterative. With each iteration $\vec{\mathbf{P}}_1$ is a better approximation to the true peak-shape vector $\vec{\mathbf{f}}$, which increases the accuracy of estimating the vectors in $\mathbf{F}$.

Correcting for linewidth variations by multiplying with negative or positive exponentials, unlike the phase and frequency corrections, changes the noise in each spectrum and thus significantly affects the overall error in the other estimations. For this reason we do not attempt to correct variations in linewidth until all other variations have been removed. In the following section some SNR considerations for this special case are presented.

In (*1*) we showed that when a single signal-related PC is present in the data the standard deviation, σ_{err}, of the difference between the estimated and true area, is given by

$$\sigma_{\mathrm{errt}} = \left(\sum_{j=1}^{m} P_{1j} \right) \sigma, \qquad [11]$$

where σ is the standard deviation of the noise in the spectra (assumed to be white and the same in every spectrum).

We assume that the existing linewidth variations in the dataset are small in order that only two signal-related PCs can sufficiently describe the data. If corrections for linewidth variations have been performed, then the standard deviation of the noise in the kth spectrum in the dataset, σ_k, after correction with $\mathbf{L}_k$ Hz lorentzian filter, is given by

$$\sigma_k = \sqrt{\frac{\mathbf{L}_k}{2m\Delta t}\left(1 - e^{-\frac{2m\Delta t}{\mathbf{L}_k}}\right)}\sigma. \qquad [12]$$

Both σ_k and σ can be easily estimated in the spectral region of the data containing only noise. In most cases σ_k is close to σ since the necessary linewidth corrections are typically small. However, even a small positive exponential can have a significant effect on the noise level. Thus if the goal of the analysis is obtaining accurate area estimation in the presence of linewidth variations, the sum of the weighted scores of the first and second PC (analogously to Eq. [1]) provides a more accurate estimate than correcting for linewidth variations:

$$A = \left(\sum_{j=1}^{m} P_{1j} \right) S_1 + \left(\sum_{j=1}^{m} P_{2j} \right) S_2. \qquad [13]$$

The standard deviation of the difference between the estimated and true area in this case is

$$\sigma_{\mathrm{est}} = \sqrt{\left(\sum_{j=1}^{m} P_{1j} \right)^2 + \left(\sum_{j=1}^{m} P_{2j} \right)^2}\, \sigma. \qquad [14]$$

Derivation of Eq. [14] (as well as Eq. [11]) assumes that $\vec{\mathbf{P}}_1$ and $\vec{\mathbf{P}}_2$ do not contain noise. However, generally all the PCs contain some noise (particularly $\vec{\mathbf{P}}_2$). Thus the quantity in Eq. [14] slightly underestimates the true error. The true error will be even greater if the linewidth variations in the dataset are large and the third- and higher order PCs are signal-related. The derivation of Eq. [14] excludes the contribution of these PCs, based on the assumption for the presence of small linewidth fluctuations.

Corrections for linewidth variations are recommended when the goal of the procedure is to end up with a single signal-related PC, as may be the case prior to pattern recognition techniques. Multiplication by positive and negative exponentials is obviously a suboptimal scheme for performing these corrections, since even minimal corrections ($\sim 1/8$ of the linewidth variations) with positive exponentials in the presence of noise can be

quite detrimental to the SNR in the resultant dataset. In view of this although linewidth variations are calculated at each iterative step, they may be corrected at the last step after sufficient number of iterations (in our experience, no more than 10) when the only remaining variations are due to linewidth.

The assumption of small fluctuations in the peak lineshape and the iterative character of the proposed procedure present a significant challenge to the stability of the procedure and its convergence to the true estimates of the peak characteristics. If there were no linewidth variations then the convergence to a single PC is a robust and reliable criteria for convergence (*2*). Correcting for linewidth variations, however, may amplify the noise in the individual spectra to such an extent that all other variations will remain undetected, even if the procedure converges to a single PC.

Here we introduce alternative convergence criteria. The total variance V in the data V is

$$V = V_S + V_N, \quad [15]$$

where V_S and V_N are the signal and noise-related variance. The components of V_S are the amplitude, frequency, phase, and linewidth variations. The total noise variance in the data is

$$V_N = m \cdot \sigma^2. \quad [16]$$

Alternatively,

$$V = V_{\vec{P}_1} + V_{\vec{P}_2} + \cdots + V_{\vec{P}_m}. \quad [17]$$

During the correction procedure the contribution of V_S to V decreases, while V_N remains constant (assuming that no corrections for linewidth variations are performed). In PC space this process is equivalent to reducing the number of significant PCs. If all but amplitude variations in the data are successfully removed, then a single PC will explain the entire signal variance and the following statement will hold:

$$\mathbf{R} = \frac{V - V_{\vec{P}_1}}{V_N} \approx 1. \quad [18]$$

For the purposes of investigating the convergence of the procedure, the ratio $\mathbf{R}$ (convergence parameter) in Eq. [18] is evaluated. If $\mathbf{R} \rightarrow 1$, then it can be concluded that the estimates of the peak characteristics are accurate. Further, the subsequent iterations of the correction procedure should not change this estimate.

From the Taylor expansions in Eq. [6] it is clear that convergence of the procedure can be ensured only for small variations in frequency and linewidth. The smaller these variations, the more accurate the representation in Eq. [10]. Alternatively, the larger these variations, the more complicated the mixtures of second and third derivatives making up the first 4 PCs. These can cause erroneous frequency estimates that can shift peaks outside of the peak shape $\vec{\mathbf{P}}_1$ and cause breakdown of the model. As a safeguard, in the most recent implementation of the procedure we have added a check to make sure the peak position after correction is within the linewidth of $\vec{\mathbf{P}}_1$. Determining the $\vec{\mathbf{P}}_1$ position and, subsequently, the frequencies of the points at the half height of the $\vec{\mathbf{P}}_1$ is straightforward, particularly because typically $\vec{\mathbf{P}}_1$ is of high SNR. Further, the highest point in the spectral region in each spectrum is determined. If the estimated frequency shifts calculated during the PCA procedure will shift the peak outside the current $\vec{\mathbf{P}}_1$, then the highest point of the peak is aligned with $\vec{\mathbf{P}}_1$ for this iteration only.

The implementation the entire estimation and correction procedure for a series of NMR spectra requires a reference peak that is present in all spectra in the dataset and its behavior is related only to the experimental/instrumental artifacts. Correction factors are determined on this reference peak and then applied to the entire spectrum. Once the experimental artifacts are removed from the reference peak, the analysis focuses on the rest of the peaks. The remaining variations in the data can be linked with anatomical or biochemical changes that occur during the experiment.

RESULTS

The new PCA procedure proposed here has been tested on sets of simulated spectra, containing a single lorentzian line with varying amplitude, frequency, phase, and linewidth and on a dataset from stimulus–recovery experiment from a rat muscle, kindly supplied by Drs. Ronald A. Meyer and Anthony T. Paganini of Michigan State University. The entire procedure, including corrections for frequency and phase shifts, as well as linewidth variations, was initially implemented in the IDL programming language (RSI, Boulder, CO) and later ported to platform independent Java language. Both programs can be obtained by request from the authors.

The first dataset was generated by multiplying a lorentzian line ($m = 512, \omega_0 = 0$ Hz, $\tau_0 = 0.008$ s, $\varphi_0 = 0$, amplitude $= 0.06283$, corresponding to a peak with height of 1, linewidth of 40 Hz, and peak area $A = 16.1$ in frequency domain) with 500 uniformly distributed random numbers between 20 and 100. We added uniformly distributed random frequency shifts (± 4 Hz) and exponentially multiplied the FIDs with uniformly distributed random numbers to simulate linewidth variations (± 8 Hz) in the data. For the purpose of comparing the errors in the estimated parameters we imposed linewidth variations of the same size as the frequency shifts in radians per second. The FIDs were then transformed to the frequency domain and the resultant spectra randomly misphased by ± 10 degrees. Finally, 500 sets of Gaussian distributed white noise (mean of 0 and a variance of 1) were added to the spectra. The SNR of the peaks in the resultant dataset ranged approximately from 10 to 50 (SNR is defined as the ratio between the height of the peak and twice the standard deviation of the noise (*9*)). Let us denote this dataset as Spectral Dataset 1 (SD1).

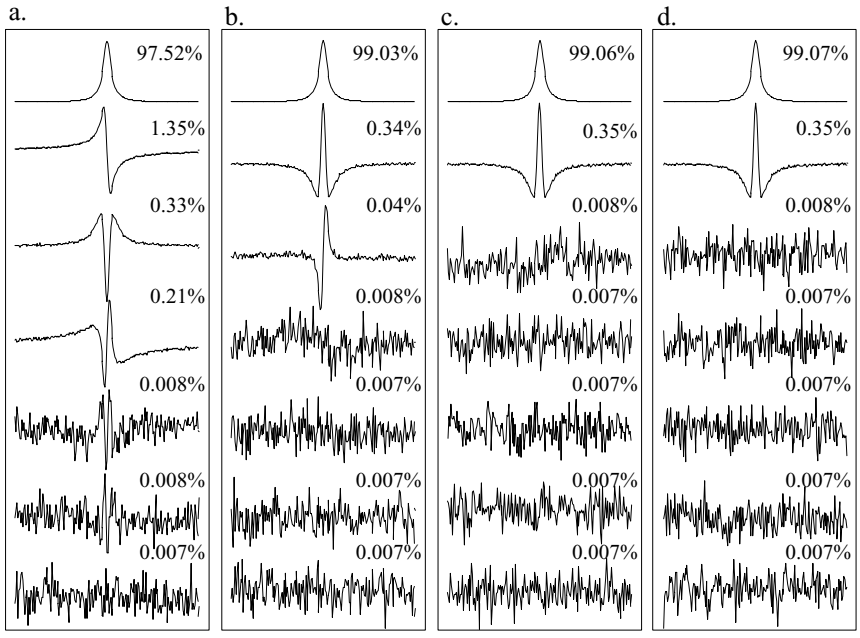

FIG. 1. First seven PCs and their corresponding normalized eigenvalues, obtained from (a) spectral data containing a single peak with varying amplitudes, frequencies, linewidths, and phases, in presence of noise; (b, c, and d) after first, second, and tenth iteration of the PCA procedure and adjusting for phase and frequency shifts variations.

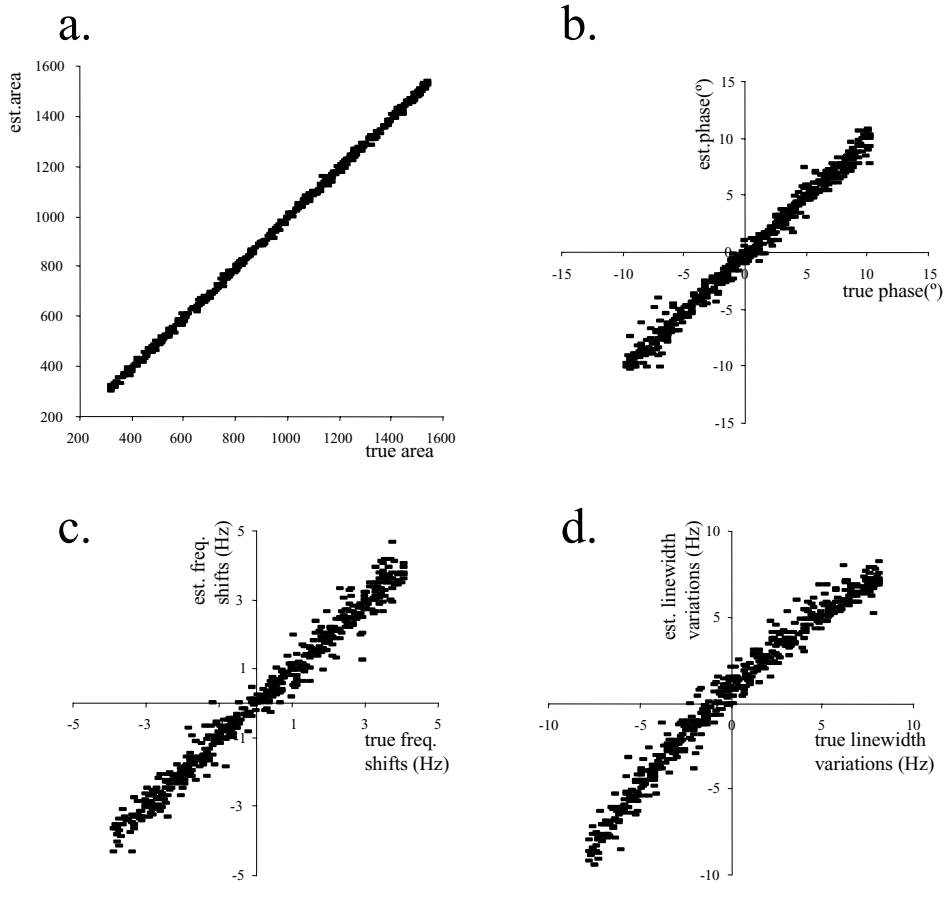

FIG. 2. Estimated versus true variations for (a) amplitude, (b) phase, (c) frequency, and (d) linewidth.

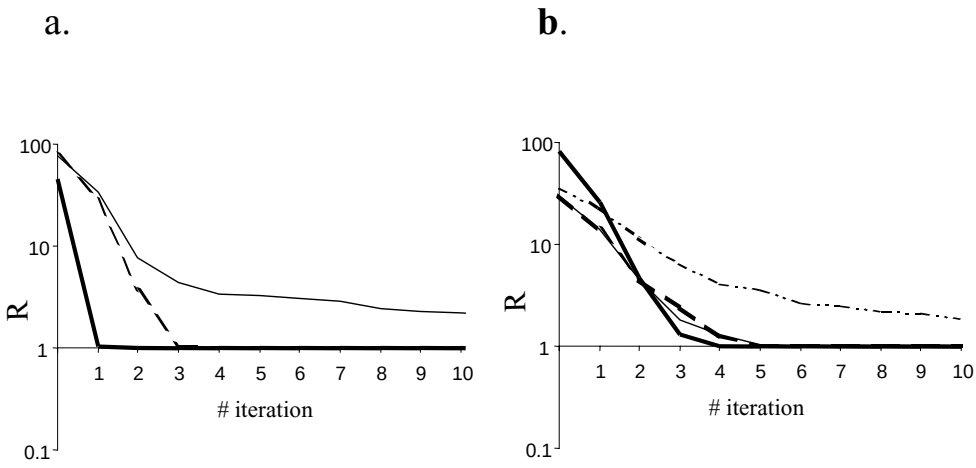

FIG. 3. Magnitude of convergence parameter **R**, displayed on log scale as a function of the first 10 iterations of the PCA correction procedure. Convergence tests were preformed on simulated data sets, described in Table 1. (a) Convergence tests on datasets with increasing phase variations: SD2 (solid line) and SD3 (thin and dashed lines). (b) Convergence tests on datasets, containing large phase and frequency variations with increasing linewidth variations: SD4 (solid line), SD5 (thin line). SD6 (dashed line), and SD7 (broken line).

To test for convergence and conditions that generated failure of the procedure six more datasets (SD2 to SD7) of 500 spectra were constructed. Table 1 summarizes the parameters of the independent sets of uniform random variations for each set. Five hundred different sets of white noise (mean of 0 and variance of 1) were added to each dataset.

For the simulated data PCA was applied to a spectral region of 201 points around the peak. The resultant first 7 PCs from SD1 are shown in Fig. 1a together with their corresponding normalized eigenvalues. The latter represent the fraction of total variance in the data that each PC explains. As mentioned above, the second and higher order PCs are mixtures of the individual shapes, related to the presence of different variations in the data.

The vectors in **F** were calculated using the first PC in Fig. 1a. The resultant offsets were calculated from Eq. [10]. The data were then corrected for phase and frequency variations. PCA was again applied, this time to the corrected dataset with the result shown in Fig. 1b. The first PC now encompasses a much larger fraction of the total variance; it is apparent that the fourth- and higher order PCs are now noise related. Another iteration yielded the PCs presented in Fig. 1c. At this point only linewidth variations are left. To demonstrate the stability of the procedure a total of 10 iterations were carried out with resultant PCs, presented in Fig. 1d. As can be seen, the signal PCs are

TABLE 1
Spectral Parameters for Generating the Simulated Spectral Datasets

	SNR	Frequency variations (Hz)	Linewidth variations (Hz)	Phase variations (°)
SD1	10–50	±4 (0.1 lw)	±8 (0.2 lw)	±10
SD2	10–50	±4 (0.1 lw)	0	±60
SD3	10–50	±4 (0.1 lw)	0	±90
SD4	10–50	±40 (2 1w)	0	±90
SD5	10–50	±40 (2 lw)	±8 (0.2 lw)	±90
SD6	10–50	±40 (2 lw)	±10 (0.25 lw)	±90
SD7	10–50	±40 (2 lw)	±20 (0.5 lw)	±90

Note. Simulated spectral datasets SD1 through SD7 consisted of 500 spectra. Uniform random distributions within the ranges, specified above were generated accordingly. A total of 500 independent distributions of white noise (mean of 0, variance of 1) were added to each of the datasets.

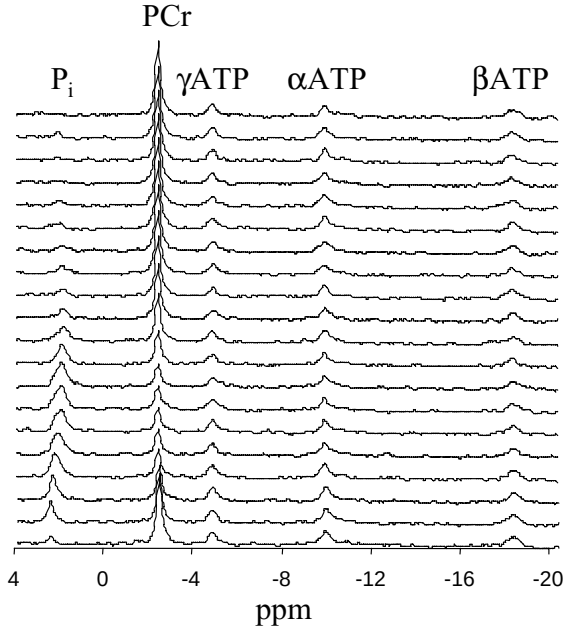

FIG. 4. ^{31}P spectra with the peaks labeled from a subset of the second experimental cycle in the rat muscle kinetic experiment. The first 10 spectra are acquired during stimulus and the last 10 during the recovery period.

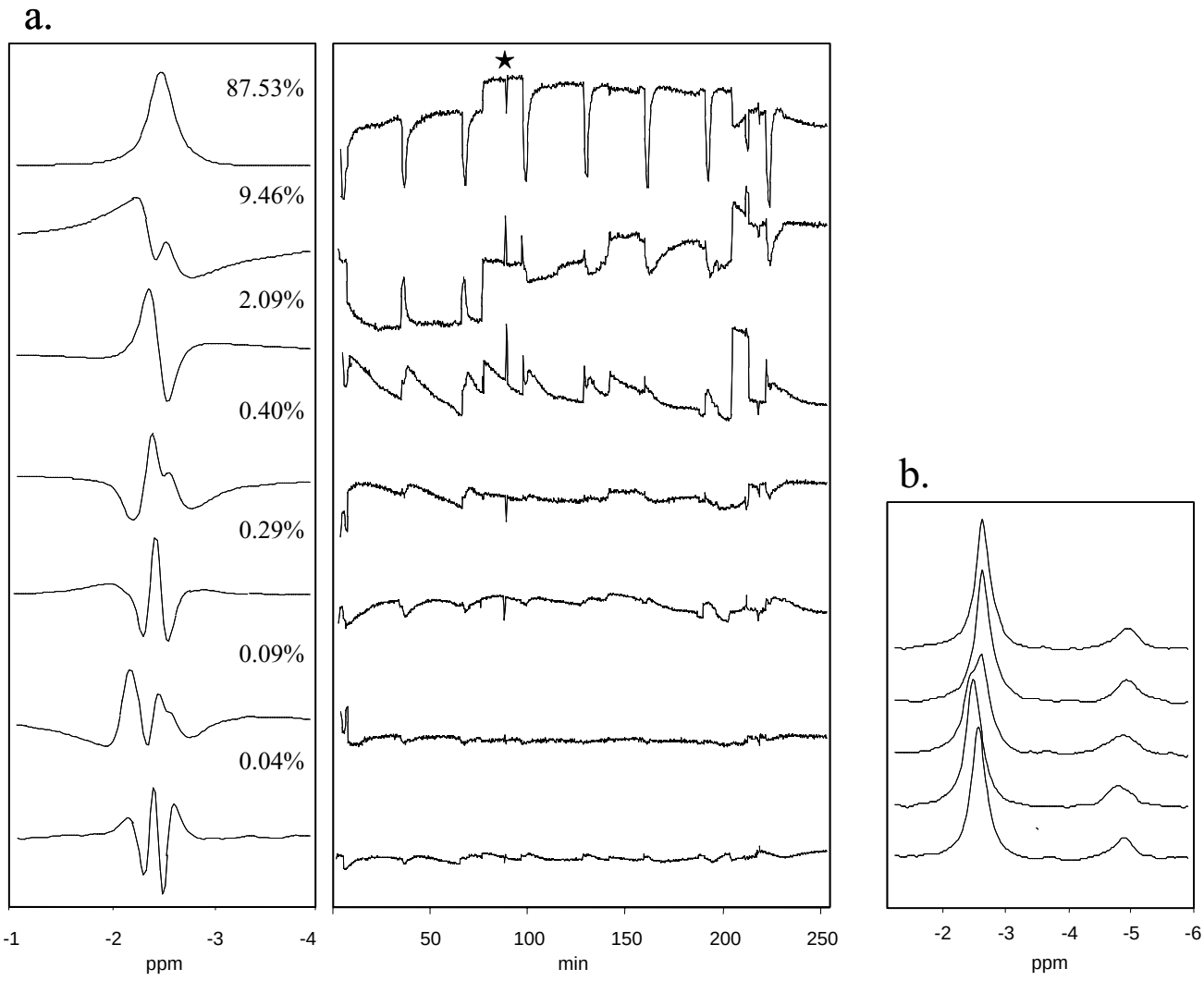

FIG. 5. (a) PCA analysis of the PCr peak in a rat muscle kinetic experiment. (Left) First seven PCs and their corresponding normalized eigenvalues. (Right) Corresponding scores, displayed on absolute scale. (b) Spectral region, containing PCr and γATP peaks in five consecutive spectra from the region marked with a star in (a).

quite stable although the higher order noise related PC vary as expected.

Figure 2 shows correlation plots between the true and estimated parameters. Since this dataset contains residual linewidth variations, the peak areas (Fig. 2a) ($R^2 = 0.9994$) were estimated using $\vec{\mathbf{P}}_1$ and $\vec{\mathbf{P}}_2$ (Eq. [13]). In the absence of linewidth variations the expected error of estimation, according to Eq. [11] involves only $\vec{\mathbf{P}}_1$ and is 5.47 in this case. For data with linewidth variations $\vec{\mathbf{P}}_2$ contributes as well and the standard deviation of the difference between the true and the estimated area should be $\sqrt{(5.47)^2 + (-5.24)^2} = 7.57$. The calculated error, however, is 8.64. As expected, the estimated error underestimates the real one slightly.

The total phase and frequency corrections (Figs. 2b and 2c) are in excellent agreement with the true values: $R^2 = 0.9867$ and 0.9791, respectively. The linewidth estimates (Fig. 2d) are in good agreement with the true ones ($R^2 = 0.9679$). However, in this correlation plot it is apparent that the best fit is not a straight line—a parabola fits the data much better ($R^2 = 0.9811$). This is due to residual second-order effects left in the data, related to the uncorrected linewidth variations.

We performed further tests for convergence, using datasets SD2 to SD7. The convergence parameter **R** from Eq. [18] is estimated prior to and after each of 10 iterations of the PCA correction procedure and presented on log scale in Fig. 3. For small frequency and phase variations (SD2) the procedure rapidly converges and maintains stability (Fig. 3a, solid line). When we attempted correcting the data in SD3 without controlling for peaks shifting outside of the first PC, the procedure did not converge well (Fig. 3a, dashed line). Alternatively, if we use the safeguard

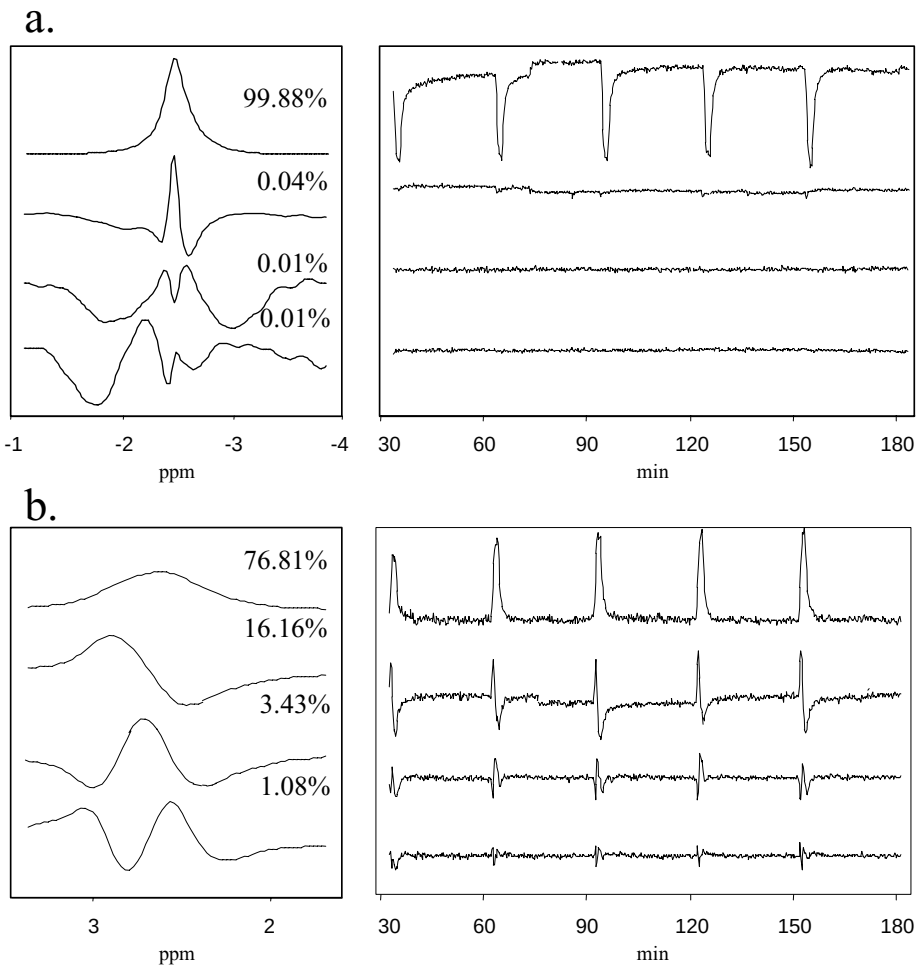

FIG. 6. (a) First four PCs from the PCr region with their corresponding normalized eigenvalues and scores, following frequency and phase adjustments. (b) First four PCs from the P_i region with their corresponding normalized eigenvalues and scores, in the corrected for PCr variations dataset.

described in the method section of only allowing shifts within the shape of the first PC, the procedure converged after the fourth iteration (Fig. 3a, fine line). This alignment procedure is used for the analysis of the datasets SD4 to SC7. Again, despite the large variations, the procedure is fully convergent for SD4 (Fig. 3b, solid line). In the presence of linewidth variations in the data (SD5 to SD7) we need to modify Eq. [18] to reflect that the second PC will also contain signal:

$$\mathbf{R} = \frac{V - V_{\vec{P}_1} - V_{\vec{P}_2}}{V_N}. \qquad [19]$$

Using Eq. [19] we determine that the procedure converges after the 5th iteration for SD5 to SD6 (Fig. 3b, thin and dashed lines), while the mixture of large linewidth, frequency, and phase variations cannot be completely overcome in the case of SD7 (Fig. 3b, broken line). We repeated the procedure 20 more iterations in this case (data not shown) and after the 15th iteration the estimate for **R** became stable, although still different from 1.

To demonstrate the procedure on real data a set of 992 ^{31}P NMR spectra (162 MHz, TR 1.0 s, 16 scans, 450 pulse, 8K sweep width, 1K data) were analyzed. These were acquired from the gastroenemius muscle of a rat (*10*) stimulated isometrically at 5 Hz for 2.1 min, followed by a 29.9-min recovery. This stimulation–recovery cycle was repeated 8 times (4.2 h). Spectra were zero-filled to 2K, linebroadened by 20 Hz lorentzian, transformed to the frequency domain, and phased uniformly. The chemical shift of phosphocreatine (PCr) was set to -2.52 ppm. Figure 4 shows a subset of the spectra from the second experimental cycle. The first 10 spectra are acquired during stimulus and the last 10 during the recovery period. It is clear that during the stimulus the PCr peak decreases and the peak of inorganic phosphate (P_i) both increases in amplitude and shifts in frequency. There are no obvious changes in the three peaks of adenosine triphosphate (ATP).

The PCr peak was selected as a reference peak since it is unaffected by changes in solution conditions under *in vivo* conditions. The PCA correction procedure was applied to 125 points

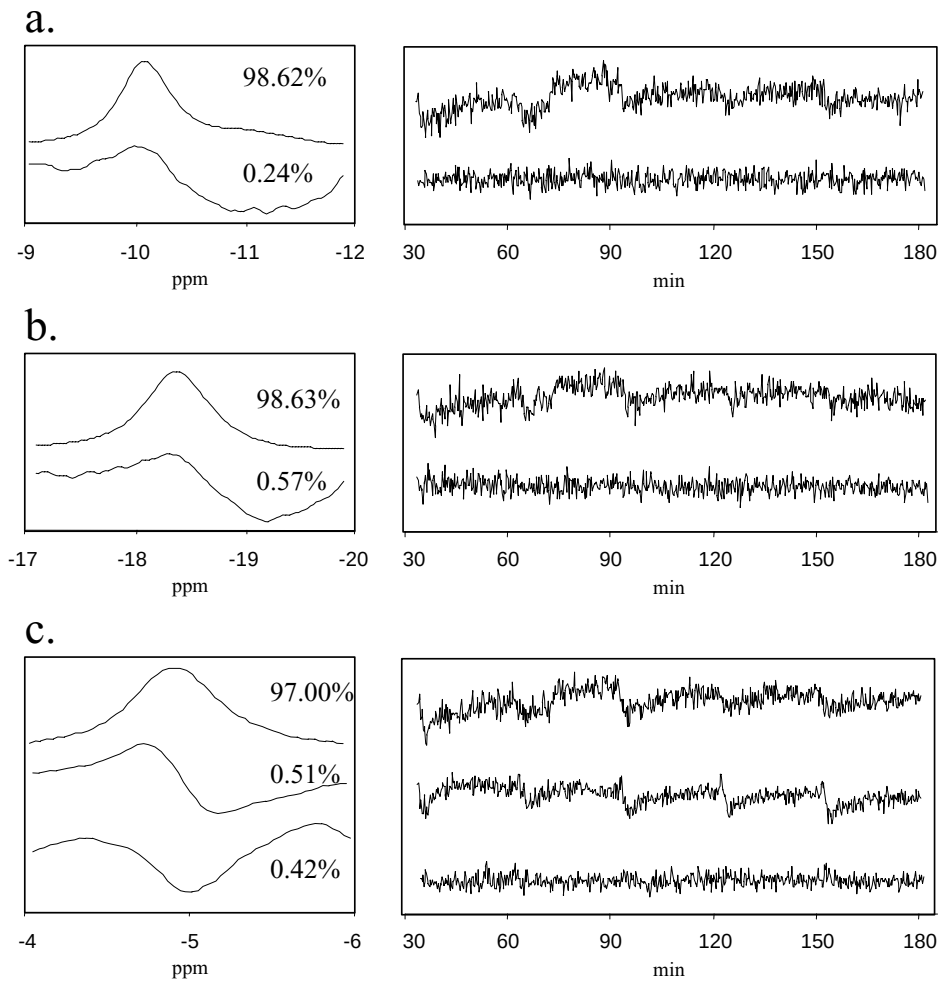

FIG. 7. PCs and their corresponding normalized eigenvalues and scores in the PCA corrected dataset from (a) αATP, (b) βATP, and (c) γATP.

around PCr [-1 ppm to -4 ppm]. The first 7 PCs (left) and their corresponding scores (right) from this analysis are presented in Fig. 5a. From the scores of the first PC, it is clear that there are 8 cycles associated with the behavior of PCr during the experiment—when the muscle is stimulated PCr decreases and when the stimulus is removed, PCr amplitude recovers back to its starting value.

Note that PCA provides a quality check of the data, since all irregularities in the cycles are detected automatically. First, there are substantial variations throughout the dataset. Second, it is apparent that the first and last two cycles of the experiment are particularly irregular and that a more stable region of exercise is exhibited in the 5 cycles in between. Finally, there are brief jumps, such as the one marked with a star in Fig. 5a. To investigate what happened at this time point, we extracted five consecutive spectra from the region around the glitch and the spectral region, containing PCr and γATP peaks presented in Fig. 5b. Note the splitting in the PCr peak in the middle spectrum, associated with a probable field jump. To continue the analysis this spectrum was replaced with the previous spectrum. We also discarded the data from the first and last two cycles. The spectra in the remaining 5 experimental cycles were frequency and phase adjusted. It took 5 iterations of the correction procedure to remove all frequency and phase variations. The largest frequency shifts applied to the data corresponded to 10 points in the spectrum ($\sim$1 linewidth of the PCr peak). The maximum and minimum phase corrections were -56° and 28°, respectively.

The phase and frequency adjusted spectra were baseline corrected (*11*) and the PCA results are shown in Fig. 6a. The eigenvalue of the first PC increased to 99.88%, indicating that following the removal of the instrumental variations, the signal in the PCr region can be adequately represented by a single peak shape. The improvement can also be seen from the scores of the first PC, showing more regular behavior of the PCr amplitudes. The shape of the second PC suggests that some linewidth variation is left in the data. Its scores (although very small in magnitude, relative to the scores of the first PC) have a regular character, related to the stimulus/recovery pattern, probably connected with

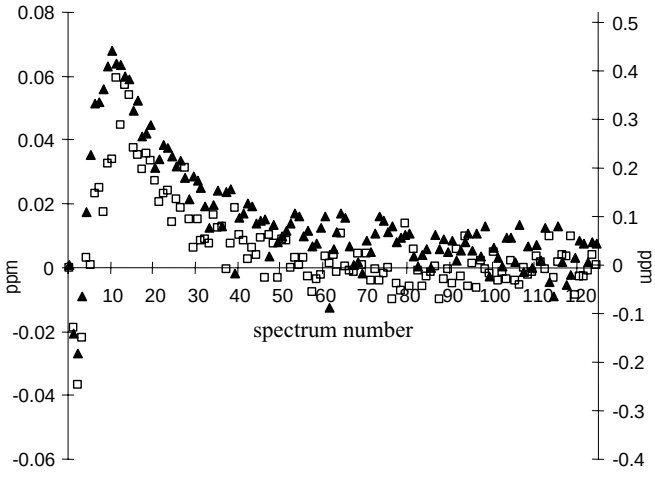

FIG. 8. Mean values of P_i (▲) and γATP (□) frequency shifts (in ppm) calculated at the same time points in the 5 exercising cycles.

the small shimming variations due to slight muscle movement in the magnet.

We next analyzed the P_i region. The PCs and their corresponding scores showed both an increase in the peak amplitude during exercise and a pH-related frequency shift (Fig. 6b).

The results of PCA analysis of the PCr-corrected data identified only one significant PC in the αATP and βATP regions (Figs. 7a, 7b), while frequency shifts were detected in the γATP positions (Fig. 7c). Note that in this analysis we observe an intensity change in the ATP level ($\sim$15% in each peak).

Since shifts of the γATP and P_i peaks are caused by pH changes, their temporal behavior should track each other over the stimulus–recovery cycle. To show this we calculated the shift of both peaks in each spectrum. Figure 8 shows average shifts at each point in the cycle calculated by averaging the five cycles together. As expected, except for scale, the results are well correlated ($\mathbf{R}^2 = 0.71$), providing excellent confirmation of the overall procedure.

CONCLUSIONS

The results presented above demonstrate the effectiveness of a PCA-based procedure for accurately quantifying peaks in large spectral datasets in the presence of simultaneous variations in amplitude, frequency, linewidth, and phase. This requires analysis in a 4-dimensional subspace derived from the peak lineshape, $\bar{\mathbf{f}}$. For its basis we choose $\bar{\mathbf{f}}$, $\frac{\partial \bar{\mathbf{f}}}{\partial \omega}$ $\frac{\partial \bar{\mathbf{f}}}{\partial \tau}$, and $\bar{\mathbf{f}}^{\mathrm{I}}$, representing the variations, rather than the basis defined by the first four PCs. Projecting Eq. [6] onto $\bar{\mathbf{f}}$, $\frac{\partial \bar{\mathbf{f}}}{\partial \omega}$, $\frac{\partial \bar{\mathbf{f}}}{\partial \tau}$, and $\bar{\mathbf{f}}^{\mathrm{I}}$ affects the efficiency of the procedure on two levels. First, this projection automatically recovers the amplitudes of the data along and only along the directions we are seeking to quantify. Second, by using vectors defined by only the first PC ($\bar{\mathbf{f}}$), rather than the second, third and fourth PCs we avoid errors introduced from the fact that as the corrections are applied these higher PCs become more and more dominated by noise. A direct consequence of this is that if a given type of variation, for example, frequency shifts, is not present in the spectral data, the data component along $\frac{\partial \bar{\mathbf{f}}}{\partial \omega}$ will be minimal, noise-related, and the subsequent correction for frequency shifts variations will not affect the data. This is the reason underlying the stability and convergence of the present procedure.

In general, by performing the parameter estimation and corrections in an iterative manner, the procedure can estimate and correct quite substantial variations. In particular, there are no limits of the magnitude of the phase variations ($\pm 90^\circ$). The procedure is also successful for quite substantial frequency shifts (± 2 linewidths). There are also no theoretical limits for linewidth variations. We have been successful recovering linewidth variations of quite large magnitude (several times the average linewidth) for simulated data in the absence of noise. The noise in the data, however, as noted above impedes the correction procedure for linewidths and thus the second order terms in Eq. [6] cannot be reduced sufficiently. Our inability to correct satisfactorily for linewidth variations is the underlying reason for the procedure failure to converge to the true value in the case of the last simulated dataset (SD7). Still, however, the procedure is stable and does not degenerate with more iterations. If the goal of the analysis is accurate linewidth estimations or reducing the spectral shapes in a given spectral region to one (as it may be for the purposes of classification) the corrections can be carried out using only negative exponentials and broadening all lines in the dataset.

The iterative manner of the proposed estimation and correction did not obstruct the procedure. It is acceptably fast (<10 s, running on Compaq UNIX Alphastation 600 MHz) and as emphasized above very stable.

Elliot *et al.* (*4*) and Wang *et al.* (*6*) have shown that applying PCA to both the real and imaginary parts of the signal, as expected, improves the accuracy of quantitation of the peak parameters by a factor of $\sqrt{2}$. Future development of PCA to spectral analysis will extend the procedure to the complex domain to include frequency and linewidth variations. The analytical considerations under Theory of this paper should serve as a theoretical basis for this implementation.

APPENDIX: ALGORITHM DESCRIPTION

1. Read entire spectral width of complex spectra data in frequency domain $C(i, j)$.

   $i = 0, \ldots, n-1$ (n– total number of spectra in C)
   $j = 0, \ldots, k-1$ (k–total number of complex points in each spectrum).

2. Define spectral peak region to be analyzed: k_f to k_l, $k_f < k_l$, $[k_f, k_l] \in [0, k-1]$

$$m = k_f - k_l + 1.$$

3. Construct real data matrix $D(i, j)$, where

$i = 0, \ldots, n-1$ ($n-$total number of spectra in C)
$j = 0, \ldots, m-1$ ($m-$total number of points in the selected peak region).

4. Principal component analysis of D:
   4.1. Calculate D's covariance matrix (Eq. [3]).
   4.2. Calculate eigenvectors $\mathbf{Q}$ and eigenvalues Λ of D's covariance matrix. The rows in $\mathbf{Q}$ are the principal components $\vec{\mathbf{P}}_1, \vec{\mathbf{P}}_2$, ...
   4.3. Calculate the scores $\mathbf{S} = \mathbf{DP}^{\mathbf{T}}$.
5. Construct the vectors in $\mathbf{F} = [\vec{\mathbf{f}}, \frac{\partial \vec{\mathbf{f}}}{\partial \omega}, \frac{\partial \vec{\mathbf{f}}}{\partial \tau}, \vec{\mathbf{f}}^{\mathrm{I}}]$.
   5.1. $\vec{\mathbf{f}} = \vec{\mathbf{P}}_1 / \sum_{j=0}^{m-1} \mathbf{P}_{1_j}$;
   5.2. $\frac{\partial \vec{\mathbf{f}}}{\partial \omega}$, where

$$\begin{cases} \frac{\partial \vec{\mathbf{f}}}{\partial \omega}\big|_{\mathrm{j}} = \mathbf{f}_{j+1} - \mathbf{f}_j, & j = 0, \ldots, m-2 \\ \frac{\partial \vec{\mathbf{f}}}{\partial \omega}\big|_{m-1} = 0, \end{cases}$$

   5.3. Calculation of $\frac{\partial \vec{\mathbf{f}}}{\partial \tau}$:
      5.3.1. Construct a complex array g of size p, where p is the closest power of two greater than or equal to m. Set the imaginary part of g to 0 and the real part as follows:

$$\begin{cases} g_j = \mathbf{f}_j, & j = 0, \ldots, m-1 \\ g_j = 0, & j = m, \ldots, p-1. \end{cases}$$

All of the remaining steps are executed in complex variables.

      5.3.2. Calculate g' through inverse FFT of g.
      5.3.3. Zero half of the point in g', i.e., construct complex vector g'' such that:

$$\begin{cases} g''_j = g'_j, & j = 0, \ldots, \frac{p}{2} - 1 \\ g''_j = 0, & j = \frac{p}{2}, \ldots, p-1. \end{cases}$$

      5.3.4. Apply 1-Hz lorentzian filter to g'':

$g'''_j = g''_j \cdot \exp(-\pi\, j \Delta t)$, where Δt is the dwell time.

      5.3.5. Calculate $\hat{g} = 2(g''' - g'')$.
      5.3.6. FFT of $\hat{g}$ to FFT($\hat{g}$).
      5.3.7. The elements of $\frac{\partial \vec{\mathbf{f}}}{\partial \tau}$ are the first m points of the real part of the result in 5.3.6:

$$\left.\frac{\partial \vec{\mathbf{f}}}{\partial \tau}\right|_j = \mathrm{Re}(\mathrm{FFT}(\hat{g}))_j, \qquad j = 0, \ldots, m-1.$$

   5.4. Calculation of $\vec{\mathbf{f}}^{\mathrm{I}}$
      5.4.1. The first 3 steps are identical to 5.3.1 to 5.3.3.
      5.4.2. Construct a vector g''' by swapping the real and imaginary part of g'', i.e.:

$$\begin{cases} \mathrm{Re}(g''') = \mathrm{Im}(g'') \\ \mathrm{Im}(g''') = \mathrm{Re}(g''). \end{cases}$$

      5.4.3. Calculate $\hat{g} = 2g'''$.
      5.4.4. FFT of $\hat{g}$ to FFT($\hat{g}$).
      5.4.5. The elements of $\vec{\mathbf{f}}^{\mathrm{I}}$ are the first m points of the real part of the result in

$$\vec{\mathbf{f}}^{\mathrm{I}}_j = \mathrm{Re}(\mathrm{FFT}(\hat{g}))_j, \qquad j = 0, \ldots, m-1.$$

6. Calculate correction factors $W(n \times 4)$ matrix, using Eq. [9] of the manuscript (we utilized the functions for matrix multiplication and inversion in IDL and Java).

7. Correction procedure:
   7.1. Frequency shifts:
      7.1.1. Determine the frequency of the highest point of the $\vec{\mathbf{P}}_1$, $h(\vec{\mathbf{P}}_1)$.
      7.1.2. Determine the frequencies of the points at the half-height of the $\vec{\mathbf{P}}_1$: h_1 and h_2.
      7.1.3. Determine the frequencies of the highest points in the spectra in D, $h(i)$.
      7.1.4. Calculate the shift for the ith spectrum in the data

$$\begin{aligned} &\delta\omega(i) = W(i,2)/W(i,1) \\ &\delta\omega(i) = 0 \quad \text{if } W(i,1) = 0 \\ &\delta\omega(i) = h(\vec{\mathrm{P}}_1) - h(i) \quad \text{if } \delta\omega(i) + h(i) < h_1 \quad \text{or} \\ &\qquad\qquad \delta\omega(i) + h(i) > h_2. \end{aligned}$$

      7.1.5. Inverse FFT complex data matrix C.
      7.1.6. If the data were zero-filled as a preprocessing step, zero as many points in the FIDs.
      7.1.7. Calculate $C_j^{\delta\omega}(i) = C_j(i) \cdot \exp(i 2\pi j \delta\omega(i))$.
      7.1.8. FFT of $C^{\delta\omega}$ to frequency domain.
   7.2. Phase correction:
      7.2.1. Calculate the phase shift for the ith spectrum in the data:

$$\begin{aligned} &\varphi(i) = W(i,4)/W(i,1) \\ &\varphi(i) = 0 \quad \text{if } W(i,1) = 0 \\ &\varphi(i) = \varphi(i) - \pi/2 \quad \text{if } W(i,1) < 0 \quad \text{and} \quad W(i,4) < 0 \\ &\varphi(i) = \varphi(i) + \pi/2 \quad \text{if } W(i,1) < 0 \quad \text{and} \quad W(i,4) > 0. \end{aligned}$$

      7.2.2. Calculate $C_j^{\varphi}(i) = C_j^{\delta\omega}(i) \cdot \exp(i\varphi(i))$.
   7.3. Linewidth correction:[2]

[2] This step is executed on the last iteration of the correction procedure.

7.3.1. Calculate the linewidth correction for the ith spectrum in the data:

$$\delta\tau(i) = W(i,3)/W(i,1)$$
$$\delta\tau(i) = 0 \quad \text{if } W(i,1) = 0.$$

7.3.2. Inverse FFT complex data matrix C^{φ}.
7.3.3. If the data were zero-filled as a preprocessing step, zero as many points in the FIDs.
7.3.4. Calculate $C_j^{\delta\tau}(i) = C_j^{\varphi}(i) \cdot \exp(\pi\,\delta\tau(i)\,j\,\Delta t)$.
7.3.5. FFT $C^{\delta\tau}$ to frequency domain.

ACKNOWLEDGMENTS

We thank Dr. William Randall for implementing the PCA correction procedure in Java language. This work was funded under NIH Grants CA41078 and CA62556.

REFERENCES

1. R. Stoyanova, A. C. Kuesel, and T. R. Brown, Applications of principal-component analysis for NMR spectral quantitation, *J. Magn. Reson. A* **115,** 265–269 (1995).
2. T. R. Brown and R. S. Stoyanova, NMR spectral quantitation by principal-component analysis. II. Determination of frequency and phase shifts, *J. Magn. Reson. B* **112,** 32–43 (1996).
3. A. C. Kuesel, R. S. Stoyanova, N. R. Aiken, C.-W. Li, B. S. Szwergold, C. Shaller, and T. R. Brown, Quantitation of resonances in biological ^{31}P NMR spectra via principal component analysis: Potential and limitations. *NMR Biomed.* **9,** 93–104 (1996).
4. M. A. Elliot, G. A. Walter, A. Swift, K. Vandenborne, J. C. Schotland, and J. S. Leigh, Spectral quantitation by principal component analysis using complex singular value decomposition, *Magn. Reson. Med.* **41,** 450–455 (1999).
5. H. Witjes, W. J. Melssen, H. J. A. 't Zandt, M. van der Graaf, A. Heerschap, and L. M. C. Buydens, Automatic correction for phase shifts, frequency shifts, and lineshape distortions across a series of single resonance lines in large spectral data sets, *J. Magn. Reson.* **144,** 35–44 (2000).
6. Y. Wang, S. Van Huffel, E. Heyvaert, L. Vanhamme, N. Mastronardi, and P. Van Hecke, Magnetic resonance spectroscopic quantification via complex principal component analysis, *in* "Proceedings of the 2000 5th International Conference on Signal Processing, World Computer Congress, Beijing, China, 2000" (Y. Baozong and T. Xiaofang, Eds.), Vol. III, pp. 2074–2077, IEEE Press, New York (2000).
7. R. Stoyanova and T. R. Brown, NMR spectral quantitation by principal-component analysis, *NMR Biomed.* **14,** 271–277 (2001).
8. T. W. Anderson, "An Introduction to Multivariate Statistical Analysis," 2nd ed., Wiley, New York (1971).
9. R. Ernst, G. Bodenhausen, and A. Wokaun, "Principles of Nuclear Magnetic Resonance in One and Two Dimensions," Oxford Univ. Press, London, 1987.
10. R. A. Meyer, A linear model of muscle respiration explains monoexponential phosphocreatine changes, *Am. J. Physiol.* **254,** 548–553 (1988).
11. S. J. Nelson and T. R. Brown, The accuracy of quantification from ^{1}D NMR spectra using the PIQABLE algorithm, *J. Magn. Reson.* **84,** 95–109 (1989).