Presented at: ICPR-06, Hong Kong, China, August 20-24, 2006
Published in: Proceedings of 2006 IEEE International Conference on Pattern Recognition

Automatic Alignment of High-Resolution NMR Spectra Using a Bayesian Estimation Approach

Zhou Wang
Dept. of Electrical Engineering
The Univ. of Texas at Arlington
zhouwang@ieee.org

Seoung Bum Kim
Dept. of Industrial & Manufacturing Systems
Engineering, The Univ. of Texas at Arlington
sbkim@uta.edu

Abstract

Nuclear magnetic resonance (NMR) spectral analysis has recently become one of the major means for the detection and recognition of metabolic changes of disease state, physiological alteration, and natural biological variation. For the pattern recognition tasks in which two or more NMR spectra need to be compared, it is critical to properly align the spectra for the subsequent pattern recognition analysis. Previous spectral alignment methods do not consider any baseline intensity variation between the spectra and disregard the effect of noise. Here we formulate the spectra alignment problem in a Bayesian statistical framework, which allows us to simultaneously and efficiently estimate the spectral shift and the baseline intensity variation in the existence of independent additive noise. Experimental results with real high-resolution NMR spectral data from human plasma demonstrate the effectiveness and robustness of the proposed approach.

1. Introduction

Pattern recognition using NMR spectra examines dynamic and time-dependent profile of metabolic responses to pathophysiological stimuli or genetic modification in an integrated biological system [4, 5]. The resonance of molecules in the sample can be represented by their chemical shifts (ppm) and the intensity values. A set of intensity values over chemical shifts lead to a spectrum (e.g., Fig. 1).

In many applications, one wish to compare a set of spectra from different samples simultaneously. However, small variations in spectra due to instrumental and environmental instabilities may significantly affect the spectral alignment and thus can interfere with direct comparisons between samples. One such example is shown in Fig. 2, where it can be observed that even a small amount of misalignment can cause large differences if the signals are directly compared point by point. Therefore, it is crucial to align the spectra *prior to* any pattern recognition processes such as feature selection and classification.

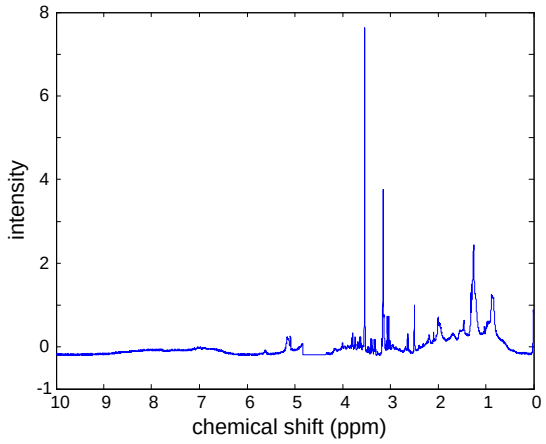

Figure 1. An NMR spectrum of human plasma obtained by a 600 MHz spectrometer.

A number of methods have been proposed to align NMR spectra. These include dynamic time warping [6], correlation optimized warping [6], partial linear fit [9], principle component analysis-based methods (for the alignment of a series of spectra) [1, 8], and genetic algorithm-based methods [2, 3]. Nevertheless, to the best of our knowledge, all these existing methods assumed that the baseline intensity variation is minimal and does not significantly affect the accuracy of alignment. In practice, however, this may not be true. As can be observed from two real NMR spectral segments shown in Fig. 2, the baseline intensity variation is a significant effect. Another fact that has often been disregarded in previous approaches is the existence of noise, which is typically observed in NMR spectral signals (see Figs. 1 and 2). Moreover, almost all existing methods involve some numerical optimization procedures that often result in increased computational complexity.

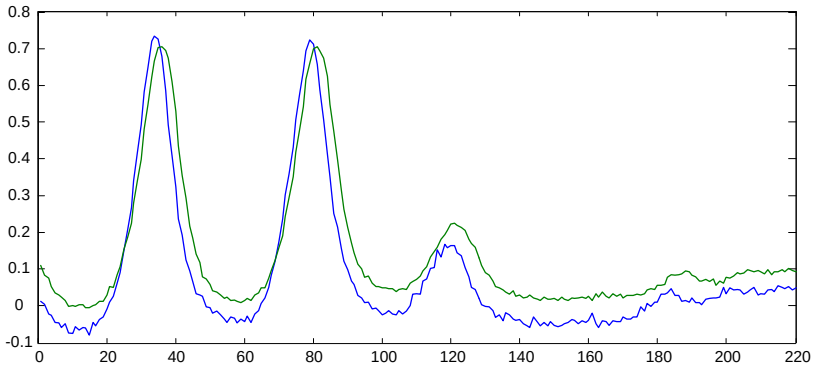

Figure 2. NMR spectra before alignment (segments extracted and enlarged from two 600 MHz NMR spectra of human plasma).

In this paper, we propose a new approach that can *simultaneously* estimate the spectral shift and the baseline intensity variation. By formulating the problem in a *Bayesian* statistical framework, the effect of noise is conveniently included. A *closed-form* solution of the problem is obtained that can be computed efficiently and shows robustness.

2. Method

2.1. Differential formulation and least square solution

Let $x(\omega)$ and $y(\omega)$ be two spectral signals to be aligned, where ω is the frequency index of the spectra. In the ideal case, the two signals represent the same spectral structure but are shifted versions of each other in both the frequency and the intensity directions. We can write

$$y(\omega) = x(\omega + \Delta\omega) + \Delta a, \tag{1}$$

where we call $\Delta\omega$ and Δa the spectral shift and the baseline intensity variation, respectively. A Taylor series expansion of the right hand side at ω_0 yields

$$y(\omega_0) = x(\omega_0) + \Delta\omega \frac{dx}{d\omega}|_{\omega_0} + \frac{(\Delta\omega)^2}{2}\frac{d^2x}{d\omega^2}|_{\omega_0} + \cdots + \Delta a\,. \tag{2}$$

In practice, the amount of the spectral shift and the baseline intensity variation are typically not fixed, but varies smoothly along the frequency axis. Therefore, Eq. (2) is only approximately true for a local spectral region-of-interest (SROI). In addition, the NMR spectral data acquired is discrete along the frequency axis. Assume that there are N discrete points within an SROI from the two signals. We denote them as $\{x(\omega_1), x(\omega_2), ..., x(\omega_N)\}$ and $\{y(\omega_1), y(\omega_2), ..., y(\omega_N)\}$, respectively. Also assume that the frequency shift $\Delta\omega$ is small, so that the second and higher order terms can be ignored. We can then write

$$\mathbf{y} = \mathbf{x} + \Delta\omega\,\mathbf{x}' + \Delta a\,\mathbf{1}, \tag{3}$$

where $\mathbf{x} = [x(\omega_1), x(\omega_2), ..., x(\omega_N)]^T$, $\mathbf{y} = [y(\omega_1), y(\omega_2), ..., y(\omega_N)]^T$, $\mathbf{x}' = [\frac{dx}{d\omega}|_{\omega_1}, \frac{dx}{d\omega}|_{\omega_2}, ..., \frac{dx}{d\omega}|_{\omega_N}]^T$, and $\mathbf{1}$ is an N-dimensional column vector with all entries equaling 1. Reorganizing Eq. (3) into a matrix operation format gives

$$\mathbf{A}\mathbf{c} = \Delta\mathbf{x}\,, \tag{4}$$

where $\mathbf{A} = [\mathbf{x}'\ \ \mathbf{1}]$, $\Delta\mathbf{x} = \mathbf{y} - \mathbf{x}$, and $\mathbf{c} = [\Delta\omega\ \ \Delta a]^T$ is a column vector containing the parameters to be estimated. The least square solution can be found by minimizing an error energy function $E(\mathbf{c}) = \|\mathbf{A}\mathbf{c} - \Delta\mathbf{x}\|^2$ and is given by

$$\hat{\mathbf{c}}_{LS} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\Delta\mathbf{x}\,. \tag{5}$$

This gives a straightforward way to simultaneously compute the best local spectral shift and baseline intensity variation parameters in the least square sense. One problem with this solution is that occasionally the matrix $(\mathbf{A}^T\mathbf{A})$ might be singular (or close to singular) and inverting the matrix may lead to unstable solutions.

2.2. Statistical modeling and Bayesian estimation

The ideal system discussed above assumed the non-existence of noise in spectral measurement, which would not be true in the real world. Motivated by the Bayesian approach in optical flow estimation [7], to account for the noise effect in a stochastic framework, we model

$$\mathbf{g} = \mathbf{A}\mathbf{c} - \Delta\mathbf{x} \tag{6}$$

as a zero-mean Gaussian random vector, in which all entries are independently and identically distributed Gaussian random variables (i.e., the noise samples are independent). The covariance matrix of $\mathbf{g}$ is thus diagonal and can be denoted as $\Lambda_n\mathbf{I}$, where Λ_n is the noise variance and $\mathbf{I}$ is the

N-dimensional identity matrix. We can then write the probability density function (PDF) of $\mathbf{g}$ for a given $\mathbf{c}$ as

$$p(\mathbf{g}|\mathbf{c}) \propto \exp\left\{-\frac{(\mathbf{A}\mathbf{c}-\Delta\mathbf{x})^T(\mathbf{A}\mathbf{c}-\Delta\mathbf{x})}{2\Lambda_n}\right\}. \quad (7)$$

Here we have used the fact that the covariance matrix is diagonal and ignored the constant in front of the Gaussian PDF (because the constant has no effect on the final solution). Based on Bayes' rule, we have

$$p(\mathbf{c}|\mathbf{g}) \propto p(\mathbf{g}|\mathbf{c})\, p(\mathbf{c}). \quad (8)$$

For the prior distribution $p(\mathbf{c})$, we model it using a zero-mean Gaussian with a diagonal covariance matrix $\mathbf{\Lambda}_p$:

$$p(\mathbf{c}) \propto \exp\left\{-\frac{1}{2}\mathbf{c}^T\mathbf{\Lambda}_p^{-1}\mathbf{c}\right\}. \quad (9)$$

By using such a prior, we have imposed that spectral shift and baseline intensity variation are uncorrelated. This is physically sensible because they are likely to be caused by independent reasons. We have also imposed a preference for small spectral shift and small baseline intensity variation. This is also a reasonable and useful assumption because large spectral shift or baseline intensity variation are unexpected, and if they do happen, it would be doubtful that they are caused by simple misalignment. It can be shown that the resulting posterior distribution is still Gaussian:

$$\begin{aligned} & p(\mathbf{c}|\mathbf{g}) \\ &\propto \exp\left\{-\frac{(\mathbf{A}\mathbf{c}-\Delta\mathbf{x})^T(\mathbf{A}\mathbf{c}-\Delta\mathbf{x})}{2\Lambda_n}\right\}\exp\left\{-\frac{1}{2}\mathbf{c}^T\mathbf{\Lambda}_p^{-1}\mathbf{c}\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\mathbf{c}^T\mathbf{\Lambda}_c^{-1}\mathbf{c} - \frac{2\,\mathbf{A}^T\Delta\mathbf{x}}{\Lambda_n}\mathbf{c} + \frac{\Delta\mathbf{x}^T\Delta\mathbf{x}}{\Lambda_n}\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\mathbf{c}-\mathbf{m}_c)^T\mathbf{\Lambda}_c^{-1}(\mathbf{c}-\mathbf{m}_c)\right\}, \end{aligned} \quad (10)$$

where

$$\mathbf{\Lambda}_c = \left(\frac{\mathbf{A}^T\mathbf{A}}{\Lambda_n} + \mathbf{\Lambda}_p^{-1}\right)^{-1} \quad \text{and} \quad \mathbf{m}_c = \mathbf{\Lambda}_c\,\frac{\mathbf{A}^T\Delta\mathbf{x}}{\Lambda_n}. \quad (11)$$

Finally, the Bayes least square (BLS) as well as the Bayes maximum a posterior (MAP) solution is given by

$$\begin{aligned} \hat{\mathbf{c}}_{BLS} = \hat{\mathbf{c}}_{MAP} = \mathbf{m}_c &= \left(\frac{\mathbf{A}^T\mathbf{A}}{\Lambda_n} + \mathbf{\Lambda}_p^{-1}\right)^{-1}\frac{\mathbf{A}^T\Delta\mathbf{x}}{\Lambda_n} \\ &= \left(\mathbf{A}^T\mathbf{A} + \Lambda_n\mathbf{\Lambda}_p^{-1}\right)^{-1}\mathbf{A}^T\Delta\mathbf{x}. \end{aligned} \quad (12)$$

$\hat{\mathbf{c}}_{BLS}$ and $\hat{\mathbf{c}}_{MAP}$ are equal because the posterior distribution is Gaussian and thus both solutions are simply the centroid of the distribution. Notice that although obtained by a different approach, this solution is consistent with the least square solution of Eq. (5) in the sense that it coincides with the least square solution when the spectral measurement is noise-free, i.e., $\Lambda_n = 0$.

The advantage of using statistical modeling and the Bayesian approach is threefold. First, the effect of noise can be well accounted for. Second, prior knowledge about the quantities being estimated (specifically, the spectral shift and the baseline intensity variation here) can be included in a natural way. Third, in the case that the matrix $(\mathbf{A}^T\mathbf{A})$ is singular or close to singular, the added diagonal matrix $\Lambda_n\mathbf{\Lambda}_p^{-1}$ in Eq. (12) makes it well-behaved, thus the solution is more robust than the least square solution.

3. Implementation and Result

Several implementation issues need to be resolved before the proposed approach is applied. First, to obtain the matrix $\mathbf{A}$, we need to compute the derivative $\mathbf{x}'$ of an input spectral signal $\mathbf{x}$. This is not trivial because of the existence of noise. A method that is often used is to apply a linear smooth filter to the signal before the differentiation operation. This is equivalent of convolving the signal with the derivative of the smooth filter. Specifically, we use a derivative of Gaussian (DoG) filter to compute $\mathbf{x}'$. Second, several parameters need to be determined, which are the noise variance Λ_n and the diagonal entries of the covariance matrix of the prior distribution $\mathbf{\Lambda_c}$. In our experiment, they were selected empirically based on the acquired data. Third, as mentioned earlier, the spectral shift and the baseline variation of intensity are approximately constant only in a relatively small SROI and may vary smoothly along the frequency axis. Therefore, we apply the Bayesian estimation approach locally within a sliding window that moves point by point across the frequency axis. This result in two sequences of estimated parameters as functions of frequency, one for local spectral shift and the other for local baseline intensity variation. Finally, to align the two spectra, we keep one of them fixed and warp the other locally based on the estimated parameters.

We tested our algorithm with a dataset of NMR spectra of human plasma obtained by a Varian INOVA 600 MHz spectrometer, where the ultimate goal is to examine metabolic perturbations induced by deficiency in sulfur amino acid. A total of 68 spectra (each with 15387 points along the frequency axis) are used in the experiment. Figs. 1 and 2 show representative spectra from the dataset. Fig. 3 is the alignment result of Fig. 2 after compensation of spectral shift only. Fig. 4 shows the result after compensation of both spectral shift and baseline intensity variation. It can be seen that the spectra are well aligned both in frequency and intensity. On average, the root mean squared error between pairs of spectra being aligned was reduced from 0.2130 (before alignment), to 0.1924 (after compensation of spectral shift only) to 0.1324 (after full alignment).

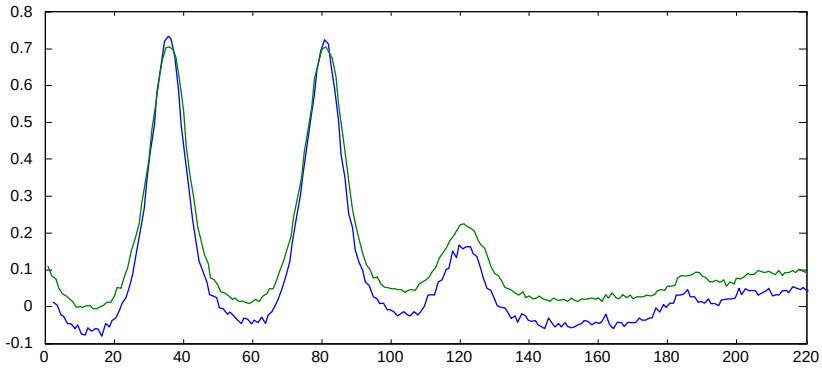

Figure 3. Aligned NMR spectra with compensation of spectral shift only.

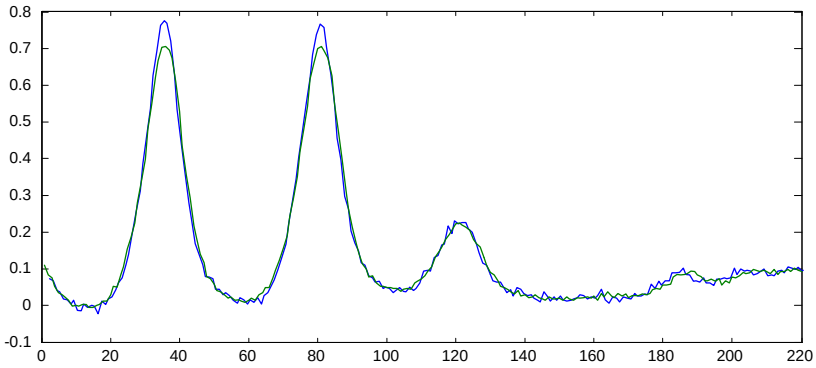

Figure 4. Aligned NMR spectra with compensation of spectral shift and baseline intensity variation.

4. Conclusion

We have proposed an algorithm for automatic alignment of NMR spectra. The novelty and advantages of our approach include 1) simultaneous estimation of both spectral shift and baseline intensity variation; 2) the use of Bayesian statistical modeling in the estimation of alignment parameters (such that noise effect can be well accounted for and prior knowledge can be included); and 3) a simple closed-form solution is obtained, leading to both efficient (compared with previous approaches that require numerical optimizations) and robust (compared with the noise-free least square solution) implementation. Experiments with real high-resolution NMR spectra of human plasma demonstrate the effectiveness of the proposed method.

References

[1] T. R. Brown and R. Stoyanova. NMR spectral quantitatioin by principal-component analysis. II. determination of frequency and phase shifts. *Journal of Magnetic Resonance*, 112:32–43, 1996.

[2] J. Forshed, I. Schuppe-Koistinen, and P. Jacobsson. Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chemica Acta*, 487:189–199, 2003.

[3] G.-C. Lee and D. Woodruff. Beam search for peak alignment of NMR signals. *Analytica Chemica Acta*, 513:413–416, 2004.

[4] J. Nicholson, J. Connelly, J. Lindon, and E. Holmes. Metabonomics: a platform for studying drug toxicity and gene function. *Nature Reviews Drug Discovery*, 1:153–161, 2002.

[5] J. Nicholson, J. Lindon, and E. Holmes. Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multi-variate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29:1181–1189, 1999.

[6] V. Pravdova, B. Walczak, and D. Massart. A comparison of two algorithms for warping of analytical signals. *Analytica Chemica Acta*, 456:77–92, 2002.

[7] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger. Probability distributions of optical flow. In *IEEE Inter. Conf. Computer Vision & Pattern Recognition*, pages 310–315, Mauii, Hawaii, June 3-6 1991.

[8] R. Stoyanova, A. W. Nicholls, J. K. Nicholson, J. C. Linidon, and T. R. Brown. Automatic alignment of individual peaks in large high-resolution spectral data sets. *Journal of Magnetic Resonance*, 170:329–335, 2004.

[9] J. T. W. E. Vogels, A. C. Tas, J. Venekamp, and J. Vander-Greef. Partial linear fit: A new NMR spectroscopy preprocessing tool for pattern recognition applications. *Journal of Chemometrics*, 10(5-6):425–438, sep-dec 1996.