Tina Memo No. 2015-010
Internal.

Fourier Domain Alignment for Independently Binned Poisson Spectra

P.D. Tar, N.A.Thacker and A. Seepujak.

Last updated
16 / 6 / 2015

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

Fourier Domain Alignment for Independently Binned Poisson Spectra

P.D. Tar and N.A.Thacker

1 Introduction

A common problem facing the analysis of spectral data is that of alignment. For example, peaks observed in ToF mass spectrometry data can shift due to slight perturbations in the hight of a sample or an initial distribution of ion velocities, depending upon the particular setup. This problem can negatively impact the analysis of data if the analysis method assumes a fixed and stable binning. If finding sources of variability in signal is important (for ICA- or PCA-like decompositions of data) then large amounts of artificial variability caused by shifting of peaks can overwhelm and conceal information of real interest. This is the case for MALDI data being used to train a Linear Poisson Model.

Various alignment methods have been proposed and are available in standard packages, such as MATLAB. Alignment methods minimise a similarity measure to achieve a "best fit" between a set of spectra and some reference spectrum. The reference spectrum may be an average computed from the spectra set, or a specific spectrum. Entire spectra can be used for alignment if there is sufficient similarity across the full range of data, or specific peaks can be identified which are common to all spectra. Common similarity measures applied include Pearson's correlation coefficient (COW, PAGA, PABS), Euclidean distances (PARS), or squared differences (PLF, DTW, PTW). Other methods including the application of Fuzzy (FW) and Bayesian (BAA) approaches. An overview can be found by Vu and Laukens 2013 for Nuclear Magnetic Resonance spectral data.

The performance of alignment methods can be assessed visually, by simply overlaying spectra before and after alignment. Preferably, a quantitative performance assessment can be achieved by comparing similarity measures. The speed of alignment methods is also sometimes considered, with optimisations being implemented in the Fourier Domain for efficient computation of correlation scores. These performance measures do not necessarily take into account the statistical properties of the data, such as changes in noise characteristics and bin correlations. The problem here is that data interpolation is often associated with data blurring which introduces correlations between adjacent data points. Summarising the success of an alignment using a single value (i.e. a similarity score) can conceal problems which may still affect processing after an apparently "successful" alignment procedure has been followed.

This document outlines an approach which is designed to align spectra consisting of independent Poisson bins. The aim is to maintain the statistical properties of the data before and after alignment. The method should achieve the best possible alignment, as measured using an appropriate similarity score, whilst simultaneously maintaining independent Poisson behaviour. These additional requirements can be tested by fitting an error model to Bland-Altman plots and computing correlation coefficients between bin pairs. The speed of the method is of little importance and will not be considered. Simple Monte Carlo spectra are used to test the method.

2 Independent Poisson Fourier Domain Alignment

This document seeks a statistically efficient alignment method for spectra composed of independent Poisson bins that outputs comparable spectra which too have independent Poisson bins. It is also important to maintain the integral of the pre-aligned spectra so as to not bias quantity measurements. Existing methods may not achieve these aims for the following reasons:

- A key reason why the statistical properties of spectra may change after alignment is the need to interpolate data points. Alignment can involve the need to shift and scale data along the x-axis (e.g. m/z) by partial bin increments, meaning that the content of bins need to be spread over new intervals using some assumed interpolation model (linear, spline etc.).
- An interpolated data point can result in correlations between adjacent data. Additionally, the noise characteristics of interpolated points may differ from the original data. For instance, if a linear interpolation is adopted then the noise on a computed point may be slightly smaller than the adjacent points from which the value was derived. For example, a data point linearly interpolated at the mid point of two measurements will have half the variance of the original data. However, if a higher-order interpolation scheme is used then instabilities may occur making the variance much larger.

- A reason why spectra may not be aligned in the most statistically efficient way is the assumptions made by the similarity measure being optimised. For example, those based upon squared distances implicitly assume Gaussian distributed bins, whereas absolute distances assume exponentially distributed bins. An efficient method will be one that matches the behaviour of the data[1].

3 A Theoretical Solution

Fourier Transforms have the property that uniform independent noise on the input data will generate uniform independent noise on Fourier coefficients. A shift in the original data can be achieved by phase shifts of these Fourier terms. Such phase shifts maintain both the magnitude and independence of noise on resulting coefficients, and when transformed back into the data space this will lead once again to independent noise on the shifted data. This process is mathematically identical to the method called 'sinc interpolation'. This process is only directly suitable for uniform independent *Gaussian* noise. But with some slight modification (using the square-root transform) we can use this property as the basis for maintaining data independence of *Poisson* distributed noise following data re-alignment.

A Fourier domain solution for spectral alignment may be possible using the following steps:

1. Select a reference spectrum containing bins, $\bar{H}_i$, where each bin, $i = 1$ to $i = N$ is a Poisson random variable.

   This may be an entire spectrum, an average of several independent spectra, or perhaps a window over a particular range. This will be the reference against which other spectra will be aligned.

2. Apply a square-root transform to give a new reference spectrum, $\bar{G}_i = \sqrt{\bar{H}_i}$.

   This step has two effects. Firstly, it transforms the Poisson data into approximately Gaussian random variables of fixed constant width. This helps to ensure uniform error characteristics are maintained in the Fourier domain. Secondly, Parseval's theorem states that the integral of a squared function is equal to its squared Fourier domain integral. If the original Poisson data is considered to be the squared function, then this means that quantities will be preserved following a shift of the data.

3. Compute the sine and cosine terms of the discrete Fourier transforms for the square-root reference spectrum giving a set of new coefficients, $\bar{a}_j$ and $\bar{b}_j$, which describe the frequency component contributions:

$$\bar{G}_i = \sum_{j=0}^{J} \bar{a}_j \sin(\frac{2\pi ij}{J}) + \bar{b}_j \cos(\frac{2\pi ij}{J})$$

   Other spectra must maximise their similarities to the reference by attempting to match these coefficients after phase-shifting each frequency component.

4. For a second spectrum requiring alignment, H_i, compute the square rooted G_i and sine/cosine coefficients, a_j and b_j. The phase of a component is then given by:

$$\phi_j = atan\frac{a_j}{b_j}$$

   such that

$$a_j = m_j \sin\phi$$

$$b_j = m_j \cos\phi$$

   where

$$m_j = \sqrt{a_j^2 + b_j^2}$$

[1] We are referring here to the distribution of a repeat measurement not the distribution of measurements. i.e. the sample noise within one bin.

5. Updated coefficients can be computed for a relative shift of δ along the original function using:

$$\theta_j = \delta * j$$

for $0 < j < \frac{n}{2}$ and

$$\theta_j = -\delta * (n - j)$$

for $\frac{n}{2} < j < n$. This assumes the index j follows the coefficient ordering produced by the FFT procedure described in Numerical Recipes in C, with the DC term at $j = 0$, which need not be shifted. Applying the phase shifts gives new coefficients:

$$a'_j = m_j \sin(\phi + \theta)$$

$$b'_j = m_j \cos(\phi + \theta)$$

6. The δ parameter is then adjusted to minimise the distance between the reference coefficients and the updated ones. As our coefficients at this point have uniform independent Gaussian noise, this can be done appropriately using a least-squares approach:

$$\arg\min_{\delta} \sum_j (\bar{a}_j - a'_j)^2 + (\bar{b}_j - b'_j)^2$$

which assumes uniform Gaussian noise on the a and b coefficients. This condition should be met, as it can be shown via error propagation that independent uniform Gaussian noise in the original domain leads to independent uniform Gaussian noise in the frequency domain.

7. Once the best fitting shift has been achieved, the new sin and cosine coefficients are used before an inverse Fourier transform giving the square-root of the aligned spectrum G'_i. Again, via error propagation, this output can be considered uniform and Gaussian. This is then converted back into Poisson data, $H'_i = G'^2_i$, by squaring.

3.1 Optimisation

To aid the optimisation of the alignment, it may be beneficial to first align to the nearest whole bin, using a simple shifting whole-bin strategy in the original (non FT) domain. A least-squares cost function on a square-root transformed spectra can be minimised:

$$f = \frac{1}{n-1} \sum_i (s\bar{G}_i - G_{i+\delta})^2$$

$$s = \frac{\sum_i G_i}{\sum_i \bar{G}_i}$$

Where δ becomes an integer shift value. In this cost function, the reference spectrum $\bar{G}_i$ is scaled to match the normalisation of the spectrum being aligned. As the reference spectrum is the sum of many spectra, it will be more accurately estimated in terms of sampling errors. f should be approximately $\frac{1}{4}$ if all spectra are repeated examples of the same spectrum, perturbed only by Poisson noise (noting that $G = \sqrt{H}$, stabilising the Poisson errors to constant value). This can be done brute-force within a set range of possible whole-bin shifts in order to find the global minimum. Partial bin shifts can then be applied in the Fourier domain using the above method, with confidence that there is only one minimum within a 1 bin shift either direction of the partially aligned spectra. A simple 1-d optimisation (such as a Golden Ratio search) might then be applied to find the final solution.

Table 1: Fit to reference spectrum and Bland-Altman error fits for aligned and non-aligned spectra

	No shift	Shift	Whole bin align	FT align
Mean fit	0.202	8.13	1.311	0.18
Fit standard deviation	0.044	9.53	0.206	0.038
B-A power fit	1.014	1.337	1.145	0.998
B-A scale fit	0.981	0.966	0.982	0.994

4 Monte Carlo results

A simple set of Monte Carlo spectra were generated, consisting of a single broad peak. These were generated initially with no shift and a Bland-Altman error model was fitted [Tina-Memo 2015-006] to show the stability of individual bins. Secondly, the spectra were shifted by a small random Gaussian amount with a standard deviation of 2 bins, giving a distribution of shifts. A Bland-Altman plot was generated again and fitted to show the increased instabilities in bin values. A whole-bin alignment was then applied to find the best integer shift which optimised cost function f. Finally, the FT alignment was applied to find the best sub-bin alignment. The results can be seen in table 1. The first and final columns show that in this simple case the FT alignment method produces aligned spectra of a similar quality (error-wise) as would be found if no misalignments were original present.

5 Summary

In summary, the problem of alignment of spectral data can be considered in two broadly different ways. Firstly, there are many alternative methods available. These different techniques can be compared directly on a dataset by dataset basis in order to find a “quickest” or “most accurate” solution, as measured using some similarity score. Alternatively, as is proposed here, a method can be designed specifically to provide particular statistical properties. Given that subsequent analysis techniques typically make assumptions regarding the properties of the spectral data (such as having independent Poisson bins) it is preferable to make choices which do not violate those assumptions. This should be the primary motivation in science application, ahead of concerns over execution time.

The use of phase shifted Fourier data to align data is simply a specific form of interpolation assumption. In this case the assumption is that any underlying data does not have spatial variations at frequencies higher than the Nyquist limit. It therefore may not be the true model of the data, resulting in additional errors in the interpolated data over and above those arising from the computational process. For arbitrary unknown data generators this may however be the best we can do.