Electronic Journal of Statistics
Vol. 8 (2014) 1748–1755
ISSN: 1935-7524
DOI: 10.1214/14-EJS900E

Analysis of proteomics data: An improved peak alignment approach[*]

Ian Zhang

Mathematics Department, Pomona College, Claremont, CA 91711

and

Xueli Liu[†]

Division of Biostatistics, City of Hope, 1500 Duarte Road, Duarte, CA 91010
e-mail: xuliu@coh.org

Abstract: Mass spectrometry (MS) data are becoming common in recent years. Prior to other statistical inferential procedures, alignment of spectra may be needed to ensure that intensities of the same protein/peptide are accurately located/identified. However, the enormous number of peaks poses challenge in handling such data. Direct applications of available curve alignment methods often do not produce satisfactory results. In this work, we propose an Automated Pairwise Piecewise Landmark Registration (APPLR) method for aligning MS data. For a pair of spectra, the most prominent peaks are given the priority to be aligned first. A weighted Gaussian kernel based similarity score is used to test warp these top peaks and spectra are then aligned according to the best match. The algorithm is implemented in an iterative way until all spectra are aligned. We illustrated the new method and two other curve alignment methods to the unlabeled total ion count data.

Keywords and phrases: Curve alignment, functional data, landmark registration, pairwise, spectrometry data, time warping.

Received August 2013.

1. Introduction

MS technology measures the mixture of proteins/peptides of biological samples to obtain relative abundance. The data are often in the format of a spectrum with a big spike as intensity associated with protein mass/charge ratios or time-of-flight in a certain range. Investigators are interested in examining spectra to identify differentially expressed proteins among groups under different conditions, e.g., between samples of healthy and diseased individuals, which may serve as potential targets of cancer therapy (Baggerly *et al.* [1]; Li *et al.* [5]; Petricoin and Liotta [6]). The unlabeled total ion count (TIC) data are an example of MS data and are introduced in Koch *et al.* [3].

[*] Main article 10.1214/14-EJS900.

[†] To whom correspondence should be addressed.

A key feature of scientific interest in MS data is the large number of peaks because they can be used to infer the existence of a particular peptide. In practice, there may be mis-alignment of such peaks among data from different labs or data generated over a long period of time in the same lab. As it can be seen, there is a rigid shift in major peaks in the unlabeled TIC data. Such misalignment must be corrected to ensure that the same protein intensities are correctly identified in a sample (Wong *et al.* [10]).

For this purpose, curve alignment/synchronization in functional data analysis may be applied. To illustrate, we applied the pairwise curve synchronization (PCS, Tang and Müller [8, 9]) and the curve alignment by moments (CAM, James [2]) to the TIC data. We noted, however, that direct application of these methods did not produce satisfactory results (see the results section). The reason may be that these methods do not take into consideration the TIC data characteristics, i.e., the extreme large number of peaks.

To address this problem, we propose an Automated Pairwise Piecewise Landmark Registration (APPLR) method that focuses on the peaks. Starting with a pair of spectra, a Gaussian kernel based similarity measure is used to determine the shift in a piecewise manner to all identified peaks between the pair. The intermediate resulting average spectrum is then warped to one of the remaining spectrum with a weight proportional to the number of spectra that are used to obtain the average spectrum. These steps iterate till all spectra get in. The procedure is applied to the TIC data. Results indicate that the performance is much better than those of PCS and CAM in recovering the underlying spectrum.

The paper is organized as follows. The underlying model and algorithm are described in section 2. Illustration to the TIC data is presented in section 3. Some discussion and concluding remarks are given in section 4.

2. The APPLR model and algorithm

2.1. The APPLR model

Assume a collection of n spectra are observed. Let (t_{ij}, f_{ij}) be the observed pairs of time points and responses for the ith sample at the jth measurement, $i = 1, \ldots, n; j = 1, \ldots, n_i$. Further, we denote by X_i the underlying continuous spectrum of the ith sample. We start with a simple time-shift model (Leng and Müller [4]) as follows:

$$f_{ij} = X(t_{ij}) + \epsilon_{ij}, X_i(t) = \mu(t - \tau_i) + \delta Z_i(t - \tau_i),$$

where $\mu(t)$ is the shared unknown shape function and Z_i are i.i.d. realizations of a stochastic process Z with mean 0 and $\mathrm{E}Z^2(t) < \infty$. The τ_i are time shifts and δ is a small positive constant. In this way, δZ_i can be viewed as a small perturbation to the shape function. And ϵ_{ij} are i.i.d. errors with mean 0 and finite variance σ_e^2.

It is apparent that this model does not require measurements at the same time points for different subjects. We partly employ this model because there are noticeable rigid shifts among spectra of samples from the same subject. In addition, a simple global shift may not work well (as will be seen in the result section) for extreme large number of peaks present in the TIC data. Therefore, we will apply the model to all prominent peaks in a piecewise manner for a pair of spectra.

Because the proposed method is based on prominent peaks (landmarks), peak quantification is a necessary pre-processing step. This involves the process of identifying peak locations on the time scales. Peaks are identified in each spectrum by computing a local maximum within some neighboring window (Yasui *et al.* [11]). Details will be described in the algorithm section.

2.2. The algorithm

In the TIC data, we have two groups of patients, responders vs non-responders. It is reasonable to align spectra within each group first. The basic idea behind the algorithm is to take two groups of curves A and B that have their peaks aligned within each group, interpolate their x-values to the same scale, average each group to generate a template curve, then warp pairs of peaks in each group to align with each other by performing linear dilations on the part of the curves between the nearest two previously aligned peaks (or the ends of the curve). If the ends are denoted by l and r, the x-value of the peak denoted by p and the target x-value by t, then this dilation can be expressed as changing the left end x-value from l to p to $l + \frac{(x-l)(t-l)}{p-l}$ and the right end x-value from r to p to $r - \frac{(r-x)(r-t)}{r-p}$. As this is done, the same linear dilation is performed on each of the curves under the template curve. The heart of the algorithm, then, only involves aligning two curves with each other, and is explained below.

After performing a baseline correction and determining the peaks using a window threshold of 1, rejecting all peaks under a certain height, the main algorithm loop starts.

1. At the n-th iteration of the loop, there are $n-1$ aligned peaks in each curve, partitioning the curve into n blocks.
2. Within each block (with edge x-coordinates of l and r, the top-ten peaks of curve A with coordinates (x_{ai}, y_{ai}) (all the peaks if less than ten) are each test-warped against all the top peaks in curve B (coordinates (x_{bj}, y_{bj})) using a Gaussian kernel based similarity score. In particular,
   a. Similarity scores for each of the top peaks are determined, by calculating which peak in the other curve matches up best with it. For peak i in curve A, this similarity score is expressed by

$$S_{ai}^{ij} = \max_j \left(\min\left(\frac{y_{ai}}{y_{bj}}, \frac{y_{bj}}{y_{ai}}\right) e^{-\left(\frac{x_{ai}^{ij}-x_{bj}^{ij}}{r-l}\right)^2 \cdot \frac{1}{\theta}} \right),$$

   where θ is a data-adaptive parameter. Data-driven methods, such as cross validation, can be used to select the θ value. In application to the TIC data, results were robust (i.e., overall shapes of the resulting mean spectra are almost invariant) over a range of θ values (e.g., $[480, 520]$). We choose $\theta = 500$ in this data set.

This similarity score factors in the ratio between the heights of the peaks and the x-distance between the peaks. The latter is determined using a Gaussian kernel that severely penalizes peaks that are too far apart on the x-axis, and is scaled by the domain of the partitioned block. We assume that the curve is relatively well-behaved, and doesn't have peaks that are significantly mismatched relative to their neighboring peaks.

b. The similarity scores are then weighted by the height of the top peaks and normalized, so that each test-warp has a maximum similarity score of 1, making the overall similarity score between peak i in curve A and peak j in curve B as follows:

$$S^{ij} = \frac{\sum\limits_{ai} y_{ai} S^{ij}_{ai} + \sum\limits_{bj} y_{bj} S^{ij}_{bj}}{\sum\limits_{ai} y_{ai} + \sum\limits_{bj} y_{bj}}.$$

c. The test-warp with the maximum similarity score among all the partitioned blocks is then actually performed on the two curves, increasing the number of aligned peaks by one.

3. After each test-warp, the x-coordinates get shifted to x^{ij}_{ai} and x^{ij}_{bj}, which is the average location of x_{ai} and x_{bj}. Other parts go through a linear dilation as described previously.

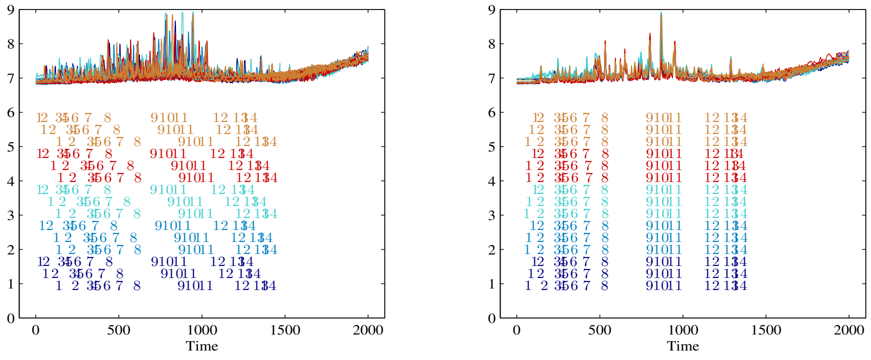
The main loop is iterated until the maximum similarity score drops below a certain minimum threshold, upon which the two curves are considered to be sufficiently well-matched.

When no such clear groupings are present, the algorithm can start with any pair of spectra, align them with each other using the above algorithm, get the average; pick any spectrum among the rest spectra, align it with the average of the pair, . . . , until all spectra get in and we then get the final average. Using the final average as the template, one can also apply the algorithm to each spectrum and the template to get the aligned individual spectrum.

We observe that the best matches often pair the top peaks of two curves along the diagonal entries, which makes intuitive sense. In addition, one does not want to do an extreme warping of the time axis which may cause unrealistic distortions.

3. Results

We illustrated the APPLR model to the TIC data. Fig. 1 (a) shows the raw (no-alignment) data with 14 markers corresponding to the key peaks in this data. These peaks locations across samples differ a lot and can be considered



(a)

(b)

Fig 1. *Data with* 14 *markers corresponding to the key peaks in the (a) raw data, i.e., no alignment, and (b) after alignment data using the APPLR model. The $x-$axis is time (in minutes) and the $y-$axis is smoothed* $\log_{10}$ *transformed intensity. The aligned curves exhibit good alignment of almost all* 14 *marked peaks.*

to be misaligned. Fig. 1 (b) shows a similar plot with the aligned curves using our method. Almost all 14 peaks are well aligned except the first and the last peaks. A closer examination shows that these two peaks are absent among some of the 15 samples. And one can more clearly discern the general pattern of the spectra based on the aligned curves.

We also carried out analysis on the TIC data using the two recently developed methods CAM (James [2]) and PCS (Tang and Müller [8, 9]). In particular, for CAM method, we used linear synchronization function, with several choices of smoothing degree freedom and penalty parameter. The result did not change much. For PCS, we tried knots number in the range of 5 to 30 and outcome did not change much either.

Although curve alignment has become an increasingly important statistical approach over the last two decades, there is no universally applicable objective measure to assess the performance of different methods in real data example where the truth is unknown. An appealing visual alternative in the literature is to look at important features in the aligned mean curve, e.g., the maximums or minimums and their locations (see also the discussion section). Table 1 summarizes these three methods in terms of the most prominent peak location and magnitude. We can see both CAM and PCS work no better than the no alignment (cross-sectional) method – the magnitude of the most prominent peak are

Table 1

Maximum magnitudes and locations for TIC data from three methods (Raw mean data also included under the No Alignment column)

Quantity	No Alignment	CAM model	PCS model	APPLR
Peak Location	108.0	103.4	98.1	106.9
Peak Magnitude	7.52	7.47	7.58	8.42

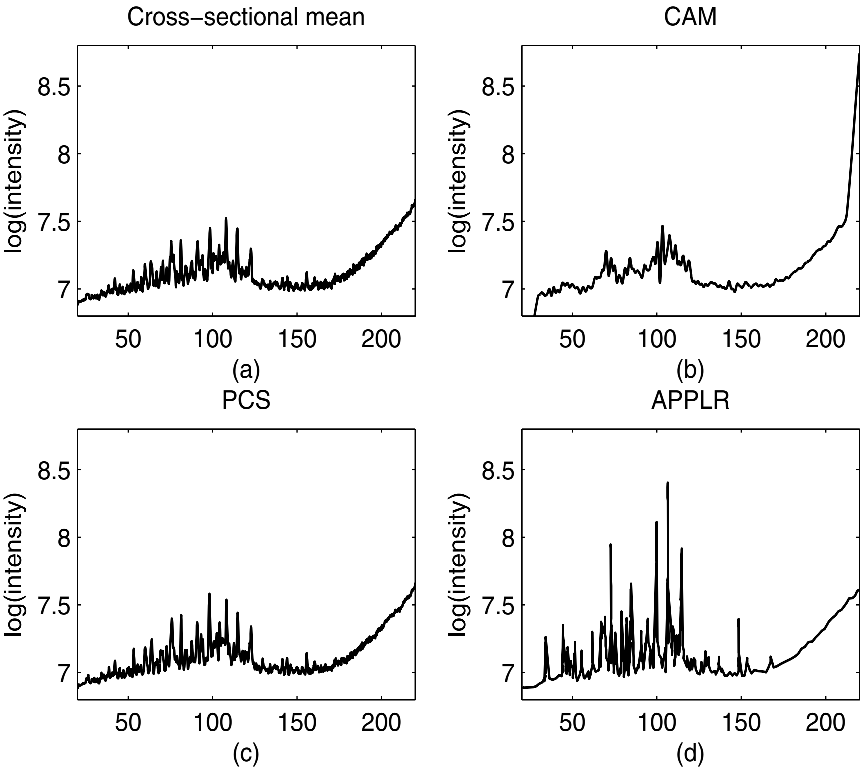


FIG 2. *Mean spectra from (a) raw data, i.e., no alignment, (b) CAM model, (c) PCS model, and (d) APPLR model. The $x-$axis is time (in minutes) and the $y-$axis is smoothed $\log_{10}$ transformed intensity. It can be seen that CAM and PCS could not recover the apparent pattern of a typical spectrum and they performed no better than the cross-sectional mean. In contrast, APPLR did very well in recovering the typical pattern of each spectrum.*

7.47 and 7.58, respectively, comparing to that of 7.52 from the no alignment method. The proposed APPLR works very well – the magnitude of the most prominent peak in the mean spectrum is 8.42.

We further investigate differences between resulting mean spectra profiles of these methods. Fig. 2 displays the mean spectra from methods of no-alignment (panel (a)), CAM (panel (b)), PCS (panel (c)), and APPLR (panel (d)), respectively. According to Fig. 2, cross sectional mean (no-alignment), CAM and PCS have similar poor performance. They could not recover the apparent pattern of each TIC spectrum, i.e., several major peaks among large number of minor peaks. Instead, major peaks were almost washed out in an averaging process. This may be due to the fact that these peaks are not well aligned through these methods because they cannot handle the extreme large number of peaks well. On the contrary, using the APPLR method, most prominent peaks are consistently identified in the resulting mean spectrum across spectra. The pattern of each spectrum is very well recovered. As a side note, the average number of peaks identified and aligned in APPLR is 46.8. Different sample spectra may have different number of peaks – the range is $[40, 52]$ and the median is 47 in the TIC data.

In summary, for the most important feature in these TIC spectra – several big spikes scattered with large number of minor peaks, the APPLR model gave meaningful mean spectrum as well as better shape by focusing on the most prominent peaks.

4. Conclusions and discussion

When mis-alignment problems are apparent in the TIC or similar MS data, one needs to make adjustment through steps of alignment so that the confidence in identifying peaks may be increased. We develop a procedure that can automatically align TIC data in a piecewise and pairwise way which makes use of information inherent in large number of peaks.

As noted, it may not be fair to compare performance of these methods by the success achieved in aligning the most prominent peak because this peak is the primary focus of APPLR whereas in the other two methods the focus is more global. In fact, similar in spirit to PCS and CAM, curve alignment methods are often designed (through the implementation of the time warping functions) to capture important features of the curves, e.g., peaks, troughs and their locations, or global characteristics, e.g., the slope of the curve over time. In particular, the Berkeley growth data (Ramsay and Silverman [7]) were used in both PCS and CAM to demonstrate that the onset times and magnitudes of the two growth spurts could be well captured. The most prominent peak in the TIC data is the most obvious landmark. One would expect these two methods will at least recover it well, if not other features.

Further, our attempt here is not necessarily to show that our approach will outperform the other two methods using some objective measure. Rather, we want to illustrate that the APPLR method can give viable result given the situation that the other two methods fail to capture the pattern when the TIC data exhibit such large number of peaks.

In the APPLR model, we implicitly assumed that the peak location variation (of the same peak across different samples) range is much smaller than the minimal distance of two adjacent peaks (different peaks). This is reasonable because generally speaking, the variation of the peak locations in the average spectrum should be relatively small. Also, Yu *et al.* [12] showed the accumulation effect of peak shift with different processing order can be ignored by using permutation experiment. This result suggests the impact of process ordering may not be very large.

Acknowledgements

The authors are very grateful for the support from the Mathematical Biosciences Institute (MBI). They are also very grateful to the referee for his/her insightful comments and suggestions.

References

[1] Baggerly, K., Morris, J., Wang, J., Gold, D., Xiao, L., and Coombes, K. (2003). A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics* **3**, 1667–1672.

[2] James, G. (2007). Curve alignment by moments. *Ann. of Appl. Stat.* **1**, 480–501. MR2415744

[3] Koch, I., Hoffman, P., and Marron, J. S. (2014). Proteomics profiles from mass spectrometry. *Electronic Journal of Statistics* **8**, 1703–1714, Special Section on Statistics of Time Warpings and Phase Variations.

[4] Leng, X. and Müller, H. (2006). Time ordering of gene co-expression. *Biostatistics* **7**, 569–584.

[5] Li, J., Zhang, Z., Rosenzweig, J., Wang, Y., and Chan, D. (2002). Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin. Chem.* **48**, 1296–1304.

[6] Petricoin, E. and Liotta, L. (2003). Mass spectrometry-based diagnostics: The upcoming revolution in disease detection. *Clin. Chem.* **49**, 533–534.

[7] Ramsay, J. and Silverman, B. (2002). *Functional Data Analysis*. New York:Springer. MR2168993

[8] Tang, R. and Müller, H. (2008). Pairwise curve synchronization for high-dimensional data. *Biometrika* **95**, 875–889. MR2461217

[9] Tang, R. and Müller, H. (2009). Time-synchronized clustering of gene expression trajectories. *Biostatistics* **10**, 32–45.

[10] Wong, J. et al. (2005). Specalign-processing and alignment of mass spectra datasets. *Bioinformatics* **21**, 2088–2090.

[11] Yasui, Y. et al. (2003). A data-analytic strategy for protein biomarker discovery: Profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* **4**, 449–463.

[12] Yu, W., Wu, B., Lin, N., Stone, K., Williams, K., and Zhao, H. (2005). Detecting and aligning peaks in mass spectrometry data with applications to MALDI. *Comp. Biol. and Chem.* **30**, 27–38.