Electronic Journal of Statistics
Vol. 8 (2014) 1742–1747
ISSN: 1935-7524
DOI: 10.1214/14-EJS900D

Analysis of proteomics data: Impact of alignment on classification*

Xiaosun Lu

Department of Statistics and Operations Research
University of North Carolina at Chapel Hill NC 27599 USA
e-mail: xiaosun@live.unc.edu

Inge Koch

School of Mathematical Sciences
The University of Adelaide SA 5005 Australia
e-mail: inge.koch@adelaide.edu.au

and

J. S. Marron

Department of Statistics and Operations Research
University of North Carolina at Chapel Hill NC 27599 USA
e-mail: marron@unc.edu

Abstract: The Fisher Rao curve registration is used for curve alignment. Quality of registration is carefully studied using zoomed views. A related linear warp is considered, and seen to give somewhat inferior performance. Alignment is also seen to give large improvements in the ultimate classification problem.

Keywords and phrases: Curve registration, distance weighted discrimination, functional data analysis.

Received August 2013.

1. Fisher Rao alignment and linear warping

The Fisher Rao domain warping approach proposed by Srivastava et al. (2011) does a good job in aligning the marked spike features in the proteomics data, presented by Koch et al. (2014). The resulting aligned functions and the warping functions are shown in the top two panels of Figure 1 respectively, colored by sample type. The numbers in the top left plot show the location of the marked spikes. The vertical ordering of the numbers is determined by the relative height of the peaks of the intensity function. It is seen that these landmarks are very well lined up after Fisher Rao alignment.

The Fisher Rao warping functions (top right in Figure 1) exhibit an approximately linear shape, especially on the interval between the two vertical dashed

*Main article 10.1214/14-EJS900.

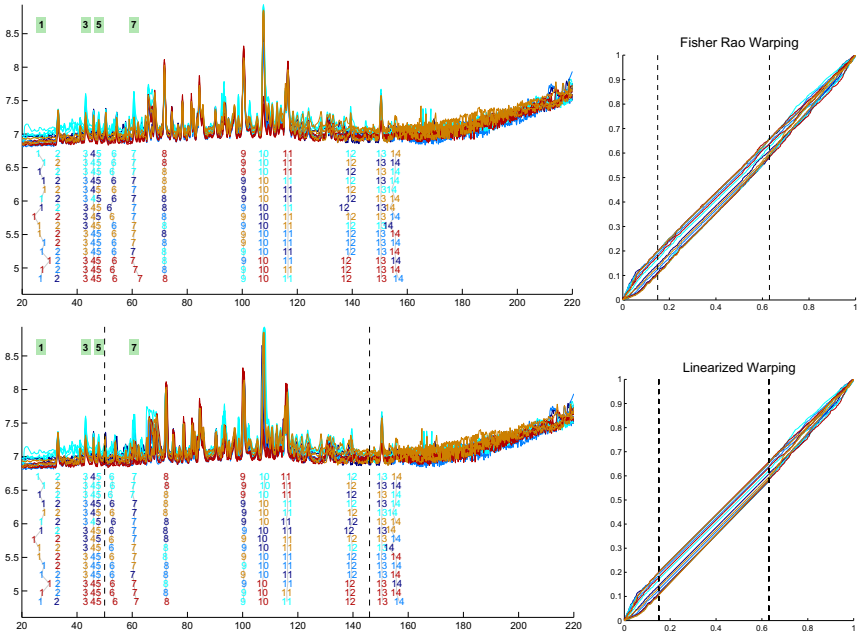

FIG 1. *Top left: Aligned functions from the Fisher Rao approach. Color represents sample type. The numbers highlight the aligned spikes. The height of the number location is consistent with the corresponding value of the function (intensity). Top right: Fisher Rao warping functions. Bottom right: The segments of the warping functions between the two vertical dashed lines are replaced by linearly fitted functions of each segment. Bottom left: Aligned functions resulting from the linearized warping functions on the right.*

lines. To investigate how well an exactly linear transformation would work for these data, we replaced the segments on this interval with linear approximations. In particular, for each original warping function, a linear regression model is fitted based on the warping function values between the two dashed lines, and the resulting linear function is used to substitute the corresponding segment of the original warping function. See the bottom right panel for the linearized warping functions. The corresponding aligned functions are shown in the bottom left panel, where both the curves and the landmarks look similar to those from the previous Fisher Rao alignment (top left). Note that the color pattern (i.e. the order of the function values) at some landmarks, such as Spikes 9 and 11, becomes slightly different after the linearized warping. This is because, computationally, these functions are discretized at limited time points. We zoomed in at each marked spike to further compare the performance of the original Fisher Rao alignment with the new alignment using the linearized warping functions. For most of the marked spikes, the original Fisher Rao alignment is better than the linearized alignment. An example is shown in the top panels in Figure 2. However, for a few marked spikes such as Spike 7 (bottom), the linearized alignment (right) may have a better performance than the original Fisher Rao alignment

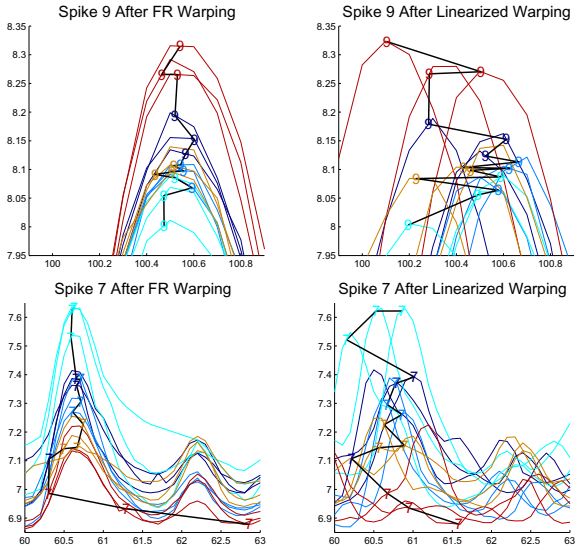

FIG 2. *Zoomed-in plots of the aligned functions at Spike 7 (bottom) and Spike 9 (top), respectively. Left: Original Fisher Rao alignment. Right: New alignment using the linearized warping functions. At Spike 9, the original Fisher Rao alignment is better, while at Spike 7 the comparison is not very clear. The Fisher Rao approach aligns Spikes 7 well for most of the samples except the red ones with low intensity, for which the linearized approach gives better results.*

(left). It is seen from the two plots that the linearized approach aligns the marked Spikes 7 of the red samples (with low intensity at Spike 7) better than the Fisher Rao approach, while for the other samples the Fisher Rao approach does a better job. This explains why, for this data set, simple linear methods, e.g. Bernardi et al. (2014), can give reasonable results, although the Fisher Rao results are slightly better.

2. Classification of responders vs. non-responders

We show that the Fisher Rao alignment greatly improves data visualization and classification of the responders to chemotherapy against the non-responders.

The PC score scatter plots before and after the Fisher Rao alignment are displayed in the left two panels in Figure 3. The symbols differentiate the responders (crosses) from the non-responders (circles). The first plot shows an overlap among different samples before alignment, while in the second plot they are better clustered (replications of the same biological sample are much closer to each other) and separated. To further investigate the difference between the responders and the non-responders, we projected the data onto the Distance-Weighted Discrimination (DWD) direction (Marron et al. (2007)) that separates these two classes. The right two panels in Figure 3 show the corresponding DWD scores before and after the alignment. It is seen that the two classes are much

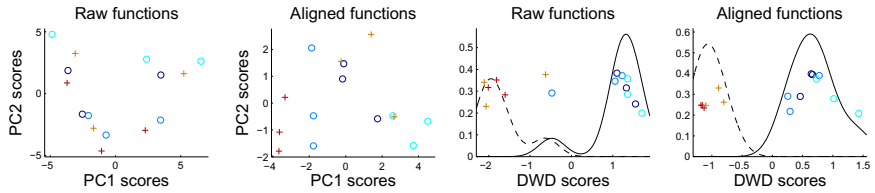

FIG 3. *The left two panels show PC score scatter plots before and after the Fisher Rao alignment respectively. The symbol indicates whether the sample responds to chemotherapy (circle for non-responders and cross for responders). The left two panels show data projections on the DWD direction (separating the responders from the non-responders) before and after the Fisher Rao alignment. The height of these points is random. The dashed and the solid curves are the densities of the two subpopulations (responders and non-responders) respectively. The aligned functions put similar cases much nearer to each other, and give better DWD separation.*

better separated after alignment, and the distribution of the two subpopulations is more Gaussian. These visual improvements brought by the Fisher Rao approach are quantitatively studied in Table 1. In particular, the clustering of the two classes was studied using the SWISS permutation tests (Cabanski et al. (2010)), and the mean difference between these two classes was studied using DiProPerm (Direction Projection Permutation) t-tests based on the DWD directions (See Wei et al. (2013) for details). The resulting p-values are listed in the table. It is seen that both data clustering and classification are greatly improved after the Fisher Rao alignment, which is consistent with the previous discussion of Figure 3. Table 1 also shows results from the linearized alignment. It is seen that the Fisher Rao warping exceeds the linearized warping in both clustering and classification of the data. Neither of these is statistically significant, but that is not surprising given the very small sample size available.

Finally, we investigate which peptides (or spikes) play an important role in classifying the responders against the non-responders. For example, in the top left panel of Figure 1, at Spike 3, the red/orange numbers (i.e. responders) are perfectly separated from the blue/cyan numbers (i.e. non-responders). That is, the reference peptide 3 is important in classification and is less prevalent in responders. Peptides 7 (small in responders) and 8 (large in responders) are also important, each with only one misclassified number. In order to identify all of the potentially important peptides, Figure 4 (top) shows the DWD loading plot based on the Fisher Rao aligned functions. The mean of these aligned functions are shown in the bottom. The green vertical lines indicate the average location of

TABLE 1
Empirical p-values in the SWISS and the DiProPerm tests before and after alignment. Alignment gives much better results of both types, especially the Fisher Rao approach

P-values	SWISS	DiProPerm
Before Alignment	0.920	0.869
After Fisher Rao Alignment	0.140	0.068
After Linearized Alignment	0.210	0.153

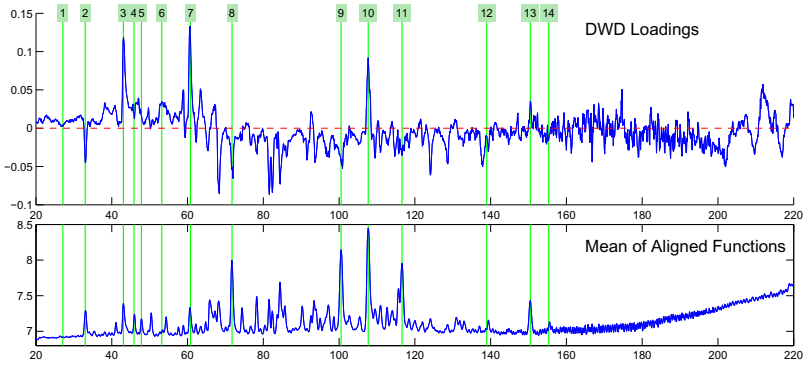

FIG 4. *Top: DWD loading curve for separating responders from non-responders based on the Fisher Rao aligned functions. The red dashed line indicates 0. Each green vertical line indicates the average location of a marked spike in the aligned functions. Bottom: Mean of the aligned functions. Peptides with a bigger absolute loading are more important in classifying responders against non-responders.*

the reference spikes in the aligned functions. Peptides with big absolute loadings are important in classifying the responders. Note that the important peptides 3, 7 and 8 discussed above correspond to prominent peaks/valleys in the loading plot. On the other hand, some reference peptides, such as 1 and 14, do not contribute much in the classification, as their loadings are close to 0. It is also seen that, some unmarked peptides turn out to be important in classifying the responders, with prominent peaks/valleys in the loading plot, such as the big negative spike between reference peptides 7 and 8. Further study of these peptides should be considered.

Acknowledgements

The authors are grateful to the Mathematical Biosciences Institute for hosting the important meeting which led to this work.

References

BERNARDI, M., SANGALLI, L. M., SECCHI, P., and VANTINI, S. (2014). Analysis of proteomics data: Block k-mean alignment. *Electronic Journal of Statistics*, 8:1714–1723, Special Section on Statistics of Time Warpings and Phase Variations.

CABANSKI, C. R., QI, Y., YIN, X., BAIR, E., HAYWARD, M. C., FAN, C., LI, J., WILKERSON, M. D., MARRON, J. S., PEROU, C. M., and HAYES, D. N. (2010). Swiss made: Standardized within class sum of squares to evaluate methodologies and dataset elements. *PLoS ONE*, 5:3.

Koch, I., Hoffman, P., and Marron, J. S. (2014). Proteomics profiles from mass spectrometry. *Electronic Journal of Statistics*, 8:1703–1713, Special Section on Statistics of Time Warpings and Phase Variations.

Marron, J. S., Todd, M. J., and Ahn, J. (2007). Distance-weighted discrimination. *Journal of the American Statistical Association*, 102(480):1267–1271. MR2412548

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. S. (2011). Registration of functional data using fisher-rao metric. arXiv:1103.3817v2.

Wei, S., Lee, C., Wichers, L., Li, G., and Marron, J. S. (2013). Direction-projection-permutation for high dimensional hypothesis tests. arXiv:1304.0796.