Electronic Journal of Statistics
Vol. 8 (2014) 1756–1758
ISSN: 1935-7524
DOI: 10.1214/14-EJS900REJ

Rejoinder: Analysis of proteomics data[*]

J. S. Marron, Inge Koch and Peter Hoffmann

Received August 2013.

The authors thank the five groups who provided excellent and interestingly different analyses of the proteomics mass spectrometry data.

In summary an encouraging observation is that all five groups achieved good quality registrations using a very diverse set of methods. An important lesson is that there is no unique best way for doing curve registration, but in fact very diverse approaches, ranging from linear transforms embedded in a block k-means structure, different versions of the Fisher-Rao approach, a Bayesian alignment approach, through a novel peak alignment approach based on the most prominent landmarks, can work quite well, even on very challenging problems such as this.

All five contributions illustrate the success of their approaches using the visual tools proposed in Section 5 of Koch *et al.* [3]. It is interesting to see that reference peptides 1 and 2 were difficult to align irrespective of the method. The alignment of reference peptides 3–11 was typically very good, but peptides 12–14 also caused problems in some of the approaches.

We are grateful to Bernardi *et al.* [1] for several interesting contributions. One remarkable feature is that they only considered linear transformations, compared to the richer transformation families that others considered. It is encouraging to see how well such a simple family of transformations fared in this apparently quite complicated example. A compelling feature of the analysis is the new block k-means alignment idea, which indeed fits the particular nested structure of this data set very well. This together with the earlier idea of k-means alignment provides some powerful new tools to the curve registration world.

Also very enjoyable is the analysis of Tucker *et al.* [5], who demonstrate very high quality registration, and also take a deep look at the motivating classification problem. Their LOO analysis is particularly informative, in terms of both amplitude and phase components. Our apriori idea was that the classification information was in the amplitude component, so we were surprised to see the interesting combined approach resulting in (slightly) better classification.

We also learned a lot from the analysis of Chen *et al.* [2], who did several things differently from the rest. First there was the enhanced preprocessing which targeted the noise much more aggressively than others did. We are not totally sure of the impact of this on the outer reference peptides, such as numbers 1 and 14. A major contribution here is a Bayes approach which tapped into

[*] Main article 10.1214/14-EJS900.

the SRVF idea in a novel and interesting way. It seems to give results that are visually quite comparable with the best of the others, perhaps because of the SRVF foundation. A major plus to the Bayesian approach is the posterior distribution, which gives a clear view of the amount of inherent variability for the important peaks. This led to useful interpretations as to which peaks were easier to estimate (all of which appear sensible). The 50000 iterations used suggests that computational time may be a factor to carefully consider in assessing this methodology.

Lu *et al.* [4] take yet another tack. While using the same registration method as Tucker *et al.* [5], they explored quite different aspects of the resulting decompositions. Note that while their aligned curves look similar to others, the peak locations are displayed quite differently in an ordering which gives a clear impression of which peptides are important for the motivating classification. The DWD loadings plot took a more explicit step in this important direction. This analysis also studied the issue of classification performance in a way which is complementary to that of Tucker *et al.* [5]. Finally, motivated by the apparently linear warping functions, they explored using exactly linear warping functions which connects with the approach taken by Bernardi *et al.* [1]. While the linear and more flexible methods give different answers, a preference is not so clear. The more flexible warping gives better small scale performance at most of the peaks (but the practical significance of this is not clear), while the linear method gave noticeably better performance at a few peaks.

We are also grateful to Zhang and Liu [6], who have added to this collection in several ways. First they provide some relevant references, not cited elsewhere. Second they have applied several current methods to these data, including CAM and PCS, and have showed that these methods do not function well in this case, probably because of the large number of peaks present in this data set. This is not surprising because those methods were designed with really different registration contexts in mind, but is good to see them appear in this quantitative study. Third, Zhang and Liu go on to suggest a clever new approach, called AP-PLR, which is aimed at overcoming the shortcomings of conventional methods, using an interesting hierarchical approach. They show that AP-PLR works much better for this example than either CAM or PCS.

References

[1] Bernardi, M., Sangalli, L. M., Secchi, P., 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.

[2] Chen, W., Dryden, I. L., Hitchcock, D. B., Le, H. (2014). Analysis of proteomics data: Bayesian alignment of functions. *Electronic Journal of Statistics* **8** 1734–1741, Special Section on Statistics of Time Warpings and Phase Variations.

[3] Koch, I., Hoffmann, P., 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.

[4] Lu, X., Koch, I., Marron, J. S. (2014). Analysis of proteomics data: Impact of alignment on classification. *Electronic Journal of Statistics* **8** 1742–1747, Special Section on Statistics of Time Warpings and Phase Variations.

[5] Tucker, J. D., Wu, W., Srivastava, A. (2014). Analysis of proteomics data: Phase amplitude separation using an extended Fisher-Rao metric. *Electronic Journal of Statistics* **8** 1724–1733, Special Section on Statistics of Time Warpings and Phase Variations.

[6] Zhang, I., Liu, X. (2014). Analysis of proteomics data: An improved peak alignment approach. *Electronic Journal of Statistics* **8** 1748–1755, Special Section on Statistics of Time Warpings and Phase Variations.