Electronic Journal of Statistics
Vol. 8 (2014) 1734–1741
ISSN: 1935-7524
DOI: 10.1214/14-EJS900C

Analysis of proteomics data: Bayesian alignment of functions*

Wen Cheng

Department of Statistics, University of South Carolina, Le Conte College, Columbia, SC 29208, USA. e-mail: chengwen1985@gmail.com

Ian L. Dryden†

School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK. e-mail: ian.dryden@nottingham.ac.uk

David B. Hitchcock

Department of Statistics, University of South Carolina, Le Conte College, Columbia, SC 29208, USA. e-mail: hitchcock@stat.sc.edu

and

Huiling Le

School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK. e-mail: huiling.le@nottingham.ac.uk

Abstract: A Bayesian approach to function alignment is introduced. A model is proposed in the ambient space, with a Dirichlet prior for the derivative of the warping function and a Gaussian process for the square root velocity function. Posterior inference is carried out via Markov chain Monte Carlo simulation. The methodology is applied to a dataset of mass spectrometry scans. Good alignment is obtained for most of the known proteins, with more uncertainty at either end of each scan.

Keywords and phrases: Ambient space, Dirichlet, Fisher-Rao, Gaussian process, Gibbs sampler, Markov chain Monte Carlo, quotient space, registration, warp.

Received August 2013.

1. Pre-processing and exploratory analysis

The mass spectrometry dataset is from Koch et al. (2013) and first of all we consider some further pre-processing. From Figure 1 we see that there is a non-constant baseline of intensities, and in common with other mass-spectrometry analyses we aim to subtract the baseline (e.g. see Fung and Enderwick, 2002; Browne et al., 2010). After some experimentation we fit a baseline using a cubic spline with $\lambda = 5$ which was subtracted from each curve. In addition, we carried

*Main article 10.1214/14-EJS900.

†To whom correspondence should be addressed.

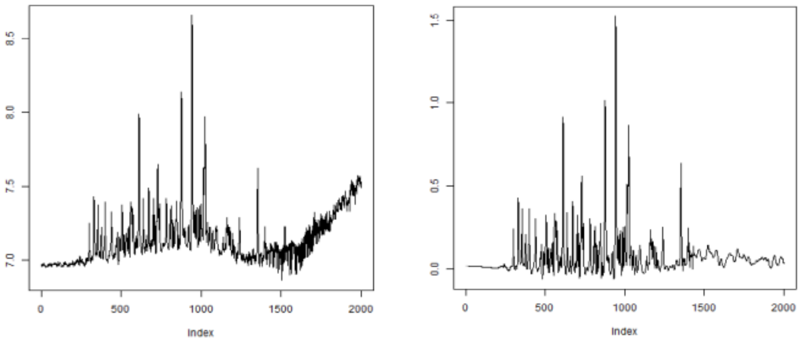

FIG 1. *Additional pre-processing of the mass-spectrometry curves. Left: A mass spectrometry scan as provided. Right: The pre-processed curve after baseline removal and additional smoothing.*

out further smoothing on the right-hand end of each curve using another cubic spline with $\lambda = 0.4$, where the aim is to remove some small undulations so that the overall baseline is reasonably flat. Of course, there is some subjectivity in these choices of pre-processing, but each curve had the same pre-processing steps carried out. Some raw data and a baseline extracted curve are given in Figure 1.

An exploratory analysis of the data makes clear from inspecting the largest peaks that a major component of the differences in the data are that the curves have been translated by different amounts in the time-axis. However, when we matched the curves using pure translations by minimizing the overall sum of square differences in intensities, we saw that the major peaks were not particularly well lined up. So we shall consider a more general transformation which makes use of the square root velocity function (SRVF) (Srivastava et al., 2011a), but we also include prior information so as to penalize transformations that are different from translations.

2. SRVF and the quotient space

We can regard each scan as a real valued differentiable curve function $f(t) : [0,1] \to \mathbb{R}^d$. The Square Root Velocity Function (SRVF) of f is then defined as $q : [0,1] \to \mathbb{R}^d$, where

$$q(t) = \frac{\dot{f}(t)}{\sqrt{\|\dot{f}(t)\|}},$$

where $\|\dot{f}(t)\|$ denotes the standard Euclidean norm of $\dot{f}(t)$ (Srivastava et al., 2011a). Here the q function is located in what is known as the ambient space (e.g., see Cheng et al., 2013). and we rescale the time in each scan so that $t \in [0,1]$. For the mass spectrometry data we have dimension $d = 1$.

Let f be warped by a re-parameterization $h \in \mathcal{H}$, i.e., $f \circ h$, where $h \in \mathcal{H} : [0,1] \to [0,1]$ is a strictly increasing differentiable warping function. The SRVF of $f \circ h$ is then given as $\tilde{q}(t) = \sqrt{\dot{h}(t)}q(h(t))$, using the chain rule. There are several reasons for using the q representation instead of directly working with original curve function f. One of the key reasons is that we would like to consider a metric which is invariant under the group G of re-parameterizations. The Fisher-Rao metric (Srivastava et al., 2011a) satisfies this desired property, i.e.,

$$d_{FR}(f_1 \circ h, f_2 \circ h) = d_{FR}(f_1, f_2),$$

although the Fisher-Rao metric is quite complicated to work with directly. However, the use of the SRVF representation simplifies the calculation of the complicated Fisher-Rao metric to an easy-to-use $\mathbb{L}^2$ metric between the SRVFs, which is attractive both theoretically and computationally. If we consider an equivalence class for a q function under G, which is denoted as $[q]$, then we have the equivalence class $[q] \in Q$, where Q is a quotient space after removing arbitrary domain warping. An elastic distance (Srivastava et al., 2011a,b) defined in Q is given as the following

$$d(q_1, q_2) = d([q_1], [q_2]) = \inf_{h \in \mathcal{H}} \|q_1 - \sqrt{\dot{h}}q_2(h)\|_2^2 = d_{FR}(f_1, f_2),$$

where $\|q\|_2$ denotes the standard $\mathbb{L}^2$-norm of q. If q_1 can be expressed as some warped version of q_2, i.e., they are in the same equivalence class, then $d([q_1], [q_2]) = 0$ in quotient space. The elastic distance is a proper distance satisfying symmetry, non-negativity and the triangle inequality (Srivastava et al., 2011a). One approach to multiple alignment is to minimize square distances in quotient space to an underlying unknown template, which is estimated using a Karcher mean (Srivastava et al., 2011b).

3. Bayesian model in ambient space

An alternative Bayesian method for multiple alignment is to consider a model in the ambient space, where the transformations are modelled with a prior probability distribution. We then remove the unwanted transformations by marginalization, and inference is based on the posterior distribution of the remaining parameters.

Note that the q-function is a continuous function in the ambient space, and so we consider a general stochastic process as a model for q. We consider a zero mean Gaussian process for the difference of two q functions, i.e., $\{q_1(t) - \tilde{q}_2(t)|h\} \sim GP$, where $\tilde{q}_2$ has been aligned relatively to q_1, and $\tilde{q}_2(t) = \sqrt{\dot{h}(t)}q_2(h(t))$ for a fixed $h(t)$. The process itself is proposed in the ambient space and we assume q_1 is untransformed while $\tilde{q}_2$ is obtained from warping q_2 to q_1. Here the alignment is carried out through the warping function $h(t)$, which contains the parameters of interest. If we use $q_1([t])$ and $\tilde{q}_2([t])$ to denote M finite points of $q_1(t)$ and $\tilde{q}_2(t)$ respectively, then the joint distribution of these M finite differences is a multivariate normal distribution based on the Gaussian process assumption, i.e,

$$\{q_1([t]) - \tilde{q}_2([t])|h\} \sim N(0_M, \Sigma_{M\times M}).$$

To simplify the problem, we assume $\Sigma_{M\times M} = \frac{1}{2\kappa} I_{M\times M}$, although more general covariance functions, such as the Gaussian or Matérn functions (Stein, 1999), could be used.

If we treat the re-parameterization function $h(t) \in \mathcal{H}$: $[0,1] \to [0,1]$ as a strictly increasing cumulative distribution function (c.d.f.), then this c.d.f. can be approximated by a set of equally spaced points along its domain $[0,1]$ and linear interpolation. Let $h([t])$ denote $\{h([t_i]), i = 0,1,2,\ldots,M\}$, the finite collection of $M+1$ discretized points and $[t_i] = \frac{i}{M}$, then we have $h([t_0]) = h(0) = 0$ and $h([t_M]) = h(1) = 1$. Further, if we let $p_i = h([t_{i+1}]) - h([t_i])$ for $i = 1,2,\ldots,M$, we have $0 < p_i < 1$ and $\sum_{i=1}^M p_i = 1$. If we denote $\boldsymbol{p}_M = (p_1, p_2, \ldots, p_M)$ and treat $\boldsymbol{p}_M$ as a random vector, we can assign a Dirichlet prior to $\boldsymbol{p}_M|h([t])$, i.e., $\pi(\boldsymbol{p}_M) \sim \text{Dirichlet}(a_1, \ldots, a_M)$. We take equal $a_i = a$ here and use $M = 40$. For $a = 1$ the prior distribution is uniform and larger values of a lead to transformations which are more concentrated on $\dot{h} = 1$ (i.e. translations). In the limit as $M \to \infty$ the warping function becomes a Dirichlet process.

The prior distribution for the concentration parameter κ is taken as a Gamma(α, β) distribution, independent of h. Combining the prior for transformation $h([t])$ and κ with the likelihood model for finite differences of two q functions, the posterior distribution for $\{h([t]), \kappa\}$ given $(q_1([t]), q_2([t]))$ is

$$\pi(h, \kappa|q_1, q_2) \propto \kappa^{p/2} e^{-\kappa\|q_1([t]) - \sqrt{\dot{h}}(q_2 \circ h)([t])\|^2} \pi(h([t]))\pi(\kappa).$$

In the above model, p represents the degrees of freedom in the model and $p = M$ in general, but if q is constrained to have unit length then $p = M-1$. In order to carry out inference on the warping function h and the concentration parameter κ, we use a Markov chain Monte Carlo (MCMC) algorithm to simulate from the joint posterior distribution. The concentration parameter κ is updated using a Gibbs sampler as the conditional posterior for κ given the other parameters is also Gamma. For $h([t])$ with $M+1$ points, a shift in $h([t_i])$ is proposed at each discrete point $(i = 1, \ldots, M-1)$ and accepted/rejected according to a Metropolis-Hastings step driven by the ratios of posterior densities. Note that $h([t_0]) = 0$ and $h([t_M]) = 1$ are both fixed. The resulting Markov chain is irreducible and aperiodic, and hence after a large number of iterations we simulated dependent values from the posterior distribution. From this Markov chain of warping functions and the corresponding registered curves, we can derive summary measures such as: (1) a cross-sectional (pointwise) mean (or median) for the registered curves; (2) a Karcher mean for the registered curves; and (3) pointwise posterior intervals for the mean of the registered curves.

If we are interested in multiple functions or curves, we can specify a mean function for the Gaussian process in the ambient space, i.e., $E(\tilde{q}_i) = \mu$, where $\tilde{q}_i = \sqrt{\dot{h}_i(t)} q_i(h_i(t))$ is a warped version of q_i through some underlying fixed h_i.

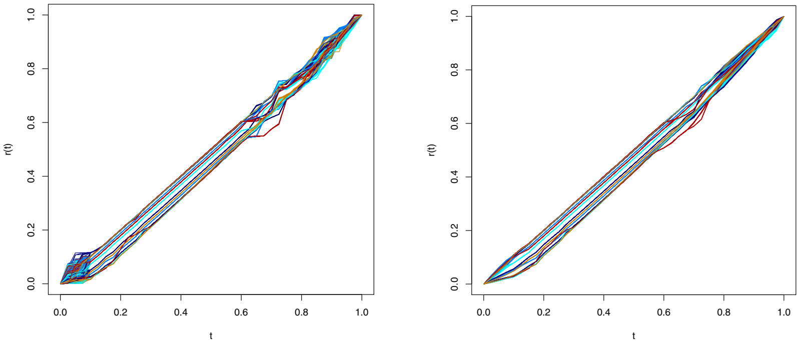

FIG 2. *Posterior warps $a = 1$ (left) Posterior warps $a = 100$ (right).*

Based on the Gaussian process assumption again, we have the random vector

$$\{\tilde{q}_i([t]) - \mu([t]) | h_i([t]), \mu([t])\} \sim N(0_M, \Sigma_{M\times M})$$

for $i = 1, 2, \ldots, n$, where n is the number of curves of interest. We take the prior distribution of μ to be a zero mean Gaussian process with large variance, independent of the other parameters. The joint posterior density for $(\mu([t]), h_1, \ldots, h_n)$ is

$$\pi(\mu, h_1, \ldots, h_n | q_1, \ldots, q_n) \propto \kappa^{np/2} e^{-\kappa \sum_{i=1}^{n} \|\mu([t]) - \tilde{q}_i([t])\|^2} \pi(\mu)\pi(h_1, \ldots, h_n)\pi(\kappa)$$

To simulate from the posterior distribution we again use an MCMC algorithm, consisting of pairwise MCMC updates from each curve to the current mean $\mu([t])$. The mean function is updated using a Gibbs step, as the conditional posterior distribution given the other parameters is also a Gaussian process.

4. Results

In Figure 2 we see plots of warping functions from the posterior distribution simulated using the Markov chain Monte Carlo algorithm. The algorithm is initalized at the quotient space estimator and then run for 50,000 iterations with a 25,000 burn-in in the final analysis.

It is clear that the posterior is much more variable for the uniform prior ($a = 1$), especially at either end of the curves. The posterior distribution of the warps is more concentrated and smoother for $a = 100$. For all the curves the posterior distribution of transformations are primarily translations of the middle part of the curve, shown by the parts of the curve which are close to a line with slope 1 but varying intercept. The strong Dirichlet prior $a = 100$ encourages $\dot{h}(t) = 1$, which seems particularly reasonable in the part with the larger peaks, as seen in both panels of Figure 2.

In Figure 3 we see the registered curves. In this plot we obtain the maximum a posteriori (MAP) estimate of each warp h_j of the jth scan to the estimated mean $\hat{\mu}$. For identifiability we also introduce the constraint constraint

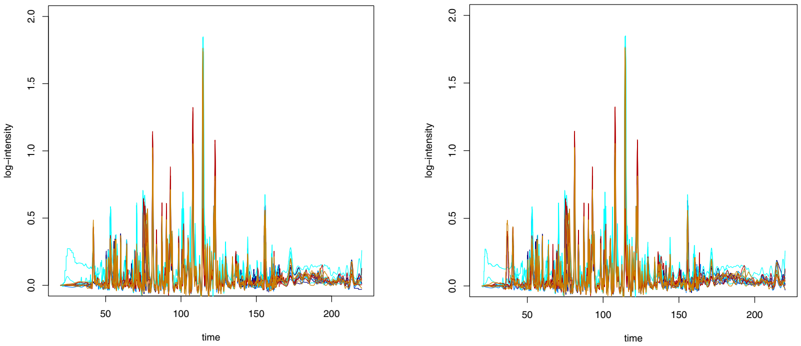

FIG 3. *Registered curves $a = 1$ (left) $a = 100$ (right). In both plots it can be seen that there is good alignment of the largest peaks, with the main differences at either end.*

$\frac{1}{n}\sum_j h_j = t$, which gives the identity tranformation as the average. We can see that in the central part of the curve with the largest peaks there is good alignment, and good agreement using both priors.

In Figure 4 we show the warped scans using prior $a \in \{1, 100, 150, 200\}$, and we also show the warped positions of the spiked proteins which are given in the answer key for each individual. The first row of integers corresponds to the warped spike positions for individual 1, replicate 1. The second row corresponds to individual 1, replicate 2, etc. In the ideal scenario of perfect alignment we should have all sets of numbers in 14 vertical columns. In this analysis the MCMC algorithm has been run for 50,000 iterations and we display every 1000th value after the burn-in period of 25,000 iterations, i.e. each number is shown 25 times. As the posterior distributions are very tight, most show very little visual variation and the more variable ones look only slightly widened. As with any MCMC scheme we need to be careful that we do not get trapped in local modes. First of all we describe $a = 100$ and we can see that spikes 3, 4, 5, 8, 9, 10, 11 have been very well aligned for all scans, and spikes 6, 7, 12, 13, 14 are mis-aligned in up to two replicates of individual 4 (in dark red) and spike 2 is mis-aligned in several individuals. Spike 1 is particularly poorly aligned, and spikes 12, 13, 14 demonstrate less accuracy in alignment in some scans compared to the rest. With a uniform prior ($a = 1$) we see that there is more variability than $a = 100$, especially at the ends, and there is less accuracy in aligning spikes 12 and 14. For the first two priors there are clear errors for individual 4 (dark red), and it can be seen in Figure 2 that the posterior warps have large kinks in this region for both choices of prior. The stronger prior at $a = 100$ gave a better alignment for higher numbered spikes than $a = 1$ in this situation. We also compared with even stronger prior $a = 150$ in Figure 4 which gave better alignment to spikes 12, and for $a = 200$ the alignment is better in spike 2 but spikes 13,14 are not as good for one of the scans as for $a = 150$. This example demonstrates the utility of a strong prior in the Bayesian procedure for this application, although the choice of prior parameters is clearly of key importance.

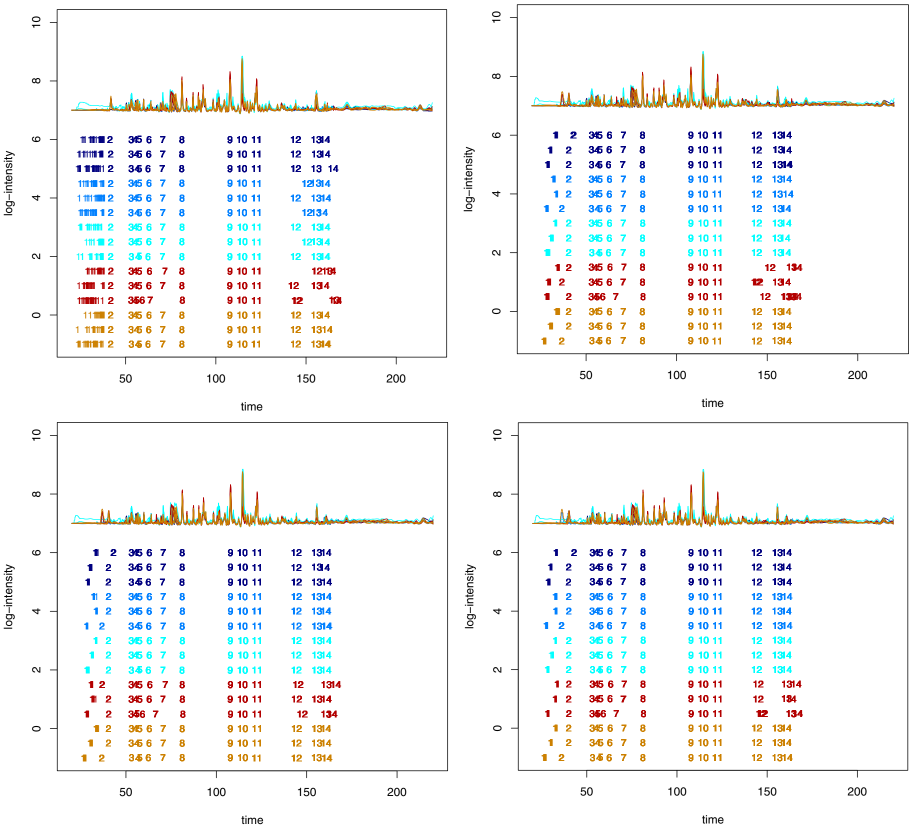

Fig 4. *The posterior distribution of the warped positions of the known spiked proteins with prior* $a = 1$ *(top left),* $a = 100$ *(top right),* $a = 150$ *(bottom left) and* $a = 200$ *(bottom right). Every 1000th value from the MCMC chain is displayed after burn-in. There is good alignment in spikes 3-11 with little posterior variability, and at either end the posterior variability is reduced with the stronger priors.*

Acknowledgment

The authors are grateful to the MBI for support. The second author acknowledges the support of a Royal Society Wolfson Research Merit Award and NSF grant 0959427.

References

Browne, W. J., Dryden, I. L., Handley, K., Mian, S., and Schadendorf, D. (2010). Mixed effect modelling of proteomic mass spectrometry data by using Gaussian mixtures. *J. R. Stat. Soc. Ser. C. Appl. Stat.*, 59(4):617–633. MR2758626

Cheng, W., Dryden, I. L., and Huang, X. (2013). Bayesian registration of functions and curves. Technical report, University of Nottingham. http://arxiv.org/abs/1311.2105.

Fung, E. T. and Enderwick, C. (2002). ProteinChip clinical proteomics: Computational challenges and solutions. *Computational Proteomics Supplement*, 32:S34–S41.

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.

Srivastava, A., Klassen, E., Joshi, S. H., and Jermyn, I. H. (2011a). Shape analysis of elastic curves in Euclidean spaces. *IEEE Trans. Pattern Anal. Mach. Intell*, 33(7):1415–1428.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E., and Marron, J. S. (2011b). Registration of functional data using the Fisher-Rao metric. Technical report, Florida State University. arXiv:1103.3817v2 [math.ST].

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York. MR1697409