Electronic Journal of Statistics
Vol. 8 (2014) 1786–1792
ISSN: 1935-7524
DOI: 10.1214/14-EJS865C

Analysis of spike train data: Classification and Bayesian alignment*

Wen Cheng

Department of Statistics, University of South Carolina, LeConte 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, LeConte 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: We analyze a data set of spike trains obtained under four different experimental conditions. We model the data curves via mixtures of normal densities. The peak locations in the fitted curves are modeled via a non-homogeneous Poisson process and classification of the spike trains into groups may be done based on the estimated spacings between peaks. We employ a Bayesian, MCMC-based registration method to align the fitted curves and summarize the data using meaningful functional statistics and posterior intervals.

Keywords and phrases: Markov chain Monte Carlo, Poisson process, registration, time warping.

Received August 2013.

1. Introduction

The spike train data in Wu et al. (2014) display the characteristic of numerous sharply peaked spikes in the functions, which is an increasingly common feature in many practical applications. The experimental data were generated by having subjects perform tasks in four different sequences (which represented four experimental conditions) and recording the response (the activity of a neuron)

*Main article 10.1214/14-EJS865.

†To whom correspondence should be addressed.

over time. It is of interest to characterize the resulting curves (spike trains) generally, and to identify differences in the spike trains across the four experimental groups. We propose to analyze the data in two stages. We will present a model for the curves, along with a method of estimating the model for each curve. This modeling itself will allow us to make some characterizations of the nature of the experimental groups. A major goal of the analysis is to register the curves so that some summary statistic (such as the cross-sectional mean curve) may be interpreted sensibly. Hence secondly, we will register the curves in the time domain to allow for straightforward summarizations and comparisons across curves.

2. Method of modeling the curves

A visual inspection of the sample data reveals that they consist of curves with humps or peaks at a series of points. Although the neural activation times are not given, we initially estimate these times by fitting a set of bumps to each curve. The peaks appear symmetric and have fairly similar width, and so a natural way to model curves with such characteristics is to use a linear combination of normal densities. Given the roughly equal hump widths, we let each Gaussian component have the same variance:

$$f(t) = \sum_{i=1}^{k} c_i g(t; \mu_i, \sigma^2),$$

where $g(t; \mu, \sigma^2)$ is the density of the $N(\mu, \sigma^2)$ distribution evaluated at t. This model enables us to estimate simultaneously both the locations (via the μ_i's) and the heights (via the c_i's) of the peaks. Browne et al. (2010) proposed a deterministic peak fitting method, which we adapt, to identify the locations of (significant) peaks. For each functional observation, there are $p = 100$ measurements made. If we rescale the time domain to $[0, 1]$, we may consider the set of peak locations to be a point process on $[0, 1]$. We sequentially fit a series of scaled Gaussian densities centered at peak locations, beginning with the largest peak height. First, we choose $i_1 \in \{1, 2, \ldots, 100\}$ such that $i_1 = \text{argmax } y_i$ and center the first Gaussian component at this point (call it μ_{i1}). To identify the correct peak height, we estimate the scale of the component density by $\hat{c}_1 = y_{i_1} / g(y_{i_1}; \mu_{i1}, \sigma^2)$, forcing the height of the density to match the maximum value of the observed curve.

The remaining peaks are found by working on a sequence of "residual curves." Let $\hat{y}_i = \hat{c}_i g(y_{i_1}; \mu_{i1}, \sigma^2)$, define the (discretized) residual curve to be $r_i = y_i - \hat{y}_i, i = 1, \ldots, p$ (and consider a normalized residual curve $r_i^* = r_i/(0.01 + y_i)$). This residual curve should be (near) zero at the location of the first peak, and should resemble the observed curve elsewhere. We center the next Gaussian component density at the location of the maximum of the residual curve, and the heights of both peaks are obtained by solving a system of two linear equations. This process of identifying peak locations continues until the residual curve is

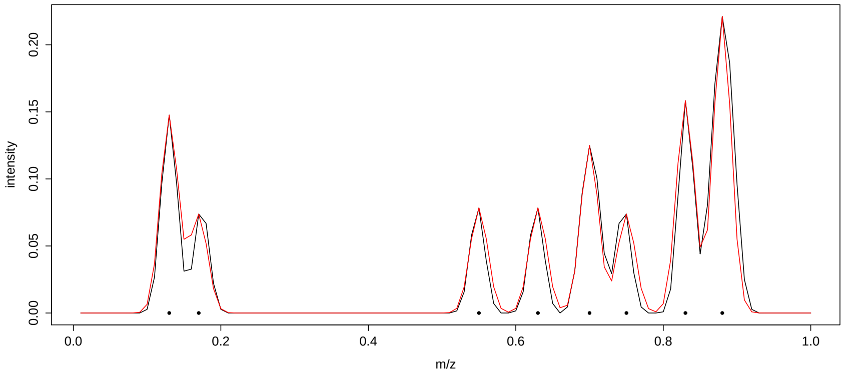

Fig 1. *Plots of the smoothed experimental data as given (black) and the fitted linear combination of Gaussian densities (red) with estimated peak locations given by black dots.*

sufficiently small, and when there are j peaks the fitted heights are obtained by solving a system of j linear equations (see Browne et al., 2010). The first function in the dataset is displayed in Figure 1 together with the fitted function and the locations of the fitted peaks.

Note that our method could be improved by fitting Gaussian peaks of the same height to each curve, as it appears this is how the curves were obtained. By using unequal peak heights, the activation rates will be underestimated in places where the activations are very close together in time. Also note that the estimation of peak locations may be unnecessary if the locations are provided with the spike train data.

3. A point process model

We model the set of peak locations for the jth curve as a non-homogeneous Poisson process with rate $\lambda_j(t)$. Then if we consider the superposition of *all* peak locations for all the spike trains, this is a non-homogeneous Poisson process with rate $\lambda(t) = \sum_{j=1}^{n} \lambda_j(t)$. The set of peak locations (ignoring group) is distributed relatively uniformly, and hence if we examine the spike trains ignoring the grouping structure, a simpler, homogeneous Poisson process would seem an appropriate model.

However, if we consider the densities of peak locations separately for the four groups, interesting distinctions emerge; see the kernel density estimates of the peak locations, by group, in Figure 2, implemented with the `bootruin` package in `R` using the method of Jones and Henderson (2007). These kernel density estimates are appropriate when the density is restricted to the domain $[0, 1]$. In Group 1, the peak location distribution has many peaks near the right end of the time axis; in Group 2, most peaks occur in the second half of the time domain. In Group 3, most peaks are in the middle of the time domain, while most peaks in Group 4 occur in the first half of the time domain.

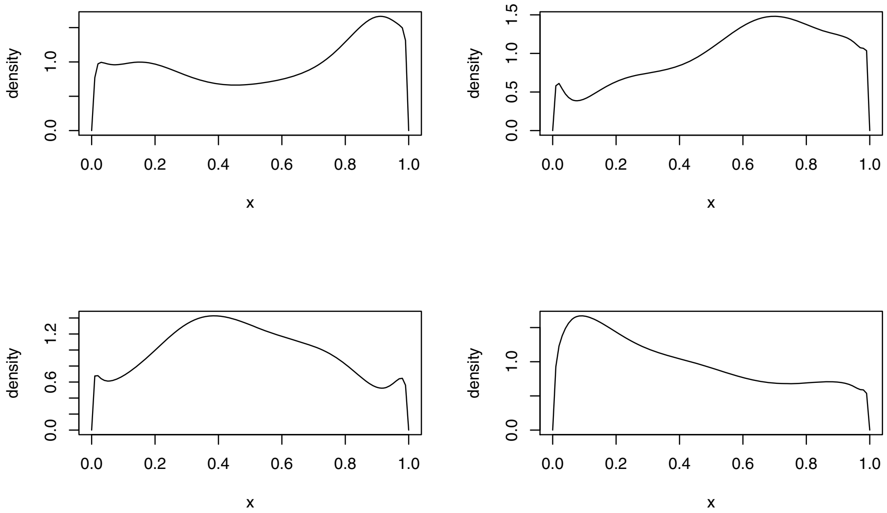

FIG 2. *Kernel density estimates of the peak locations, separate by group. Top left: Group 1. Top right: Group 2. Bottom left: Group 3. Bottom right: Group 4.*

The Poisson process model serves as a tool for an important aim of this analysis: the classification of the spike trains. Assuming this Poisson process for the peak locations, the spacings (waiting times) X_i between peaks are approximately exponential(λ_j). For each spike train (with N spacings), the log-likelihood

$$\log L \approx \sum_{i=1}^{N} \log \hat{\lambda}_i - \sum_{i=1}^{N} x_i \hat{\lambda}_i$$

can be calculated for each curve, given an estimate of λ_i. Estimating λ_i for each group, we classified the curve into the group that yields the highest log-likelihood for that curve. Applying this approach to the real spike train data set, the rate of correct classification was 78.3%. An analogously derived likelihood assuming the spacings to be independent Gaussian random variables produced a slightly higher overall classification rate, 81.7%. These classification rates should become higher if the original spike locations were available, rather than the smoothed function which was given.

4. Registration of the curves via an MCMC algorithm

Our method for registering the curves is essentially a Bayesian MCMC implementation based on the Square Root Velocity Function (SRVF) approach of Srivastava et al. (2011). The details of the approach are given in Cheng et al. (2014, Section 2) and Cheng et al. (2013).

In terms of the practical application, we registered a sample of curves with respect to a single template curve. The initial template curve was chosen by first

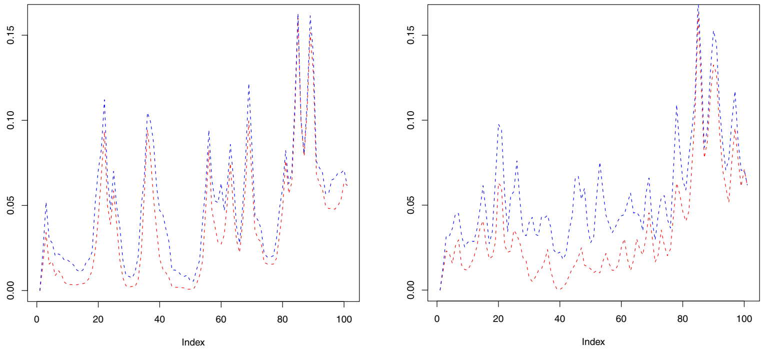

Fig 3. *The effect of the prior parameters on the precision of the posterior intervals for the Group 1 mean. Left: Posterior intervals using Dirichlet prior with $a = 1.2$. Right: Posterior intervals using Dirichlet prior with $a = 100$.*

identifying the curves (among the set of curves to be aligned) having the median number of peaks. Of that set $\mathcal{C}$ of "median-peaked" curves, we chose the curve in $\mathcal{C}$ whose vector of peak locations had the smallest Euclidean distance to the mean vector of peak locations for curves in $\mathcal{C}$. Intuitively, the initial template curve was among the most typical in terms of number of peaks, and among that set $\mathcal{C}$ was the most typical in terms of peak locations. The initial template was then updated (via a Gibbs sampler) during the MCMC process.

To allow for sensible interpretation of summary statistics (such as the cross-sectional mean curve $\bar{f}(t) = M^{-1}\sum_{i=1}^{M} f(t_i)$), we choose to register the fitted curves obtained via the model described in Section 2. The choice of parameters on the Dirichlet prior makes a substantive difference in the pointwise 95% posterior intervals for the cross-sectional mean (i.e., having 0.95 posterior probability at *each* point t of containing the mean). An overly diffuse (near uniform) prior ($a = 1.2$) results in posterior intervals that are wider at the edges of the time domain but perhaps inappropriately narrow in the middle. With a tighter prior ($a = 100$), the posterior intervals display the variability more appropriately (see Figure 3, which shows the pointwise posterior intervals for the mean curve for Group 1 under each hyperparameter specification). Intuitively, the strong $a = 100$ prior encourages the warping function velocity $\dot{h}(t) = 1$.

Some differences are apparent in the cross-sectional means of the four groups using the registered fitted curves of the sample (see Figure 4). The regions in each group where the mean curve is dense in peaks tend to match what was seen in the peak-location density estimates (compare Figures 2 and 4). In particular, note that the central curves for Groups 2, 3, and 4 have their most prominent peak activity in the late, middle, and early regions of the time domain, respectively. To some degree, this is dependent on an appropriate choice of initial template curve for each group. Recall that while we chose the initial template curve as described in Section 4, this template is updated (via a Gibbs sampler) during the MCMC process.

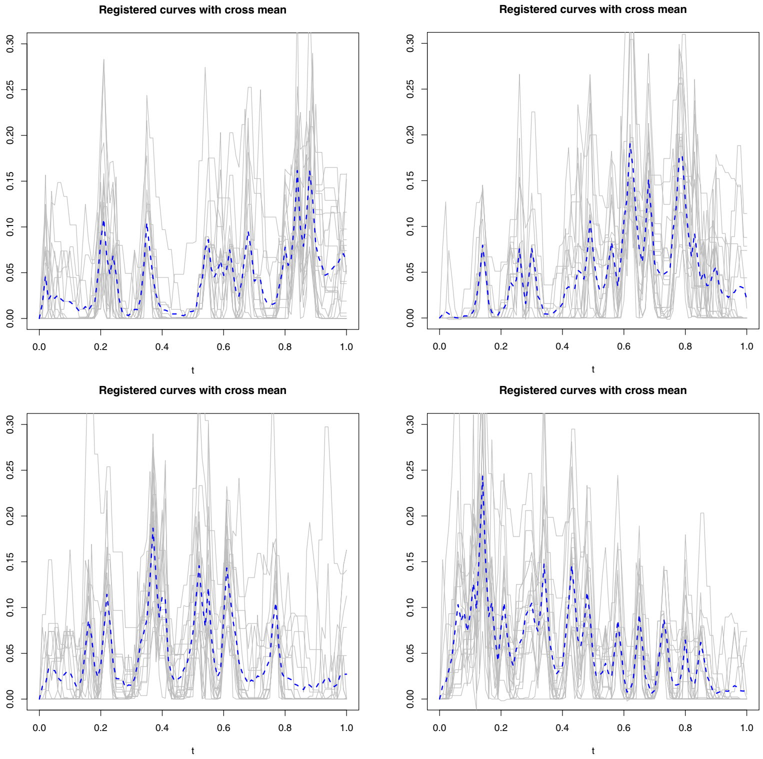

Fig 4. *Registered fitted curves, with cross-sectional mean, by group. Top left: Group 1. Top right: Group 2. Bottom left: Group 3. Bottom right: Group 4.*

If the group labels for the observed data are known (as they are for the spike trains), the cross-sectional means could aid in classifying new spike trains to one of the groups. One could register the new curves and use standard functional classification methods (see Baíllo et al., 2011). Another approach is based on the warping function $h(t)$ of a new curve: One could align the curve to each group mean (yielding warping functions $h^{[1]}(t), \ldots, h^{[4]}(t)$) and assign the curve to group j where the distance between $h^{[j]}(t)$ and the identity warp $h(t) = t$ is minimized, for $j = 1, \ldots, 4$.

5. Conclusion

Although there was a great deal of within-group variation, our approach found differences in the pattern of peak positions across the spike trains that corresponded fairly well to the four experimental groups. A likelihood-based classification method performed reasonably well in classifying individual curves to experimental groups. In addition, our MCMC approach to registration allows us to simultaneously register the curves and compute functional summary statistics. We also obtained posterior intervals that reflected the uncertainty in both the registration process and the mean curve.

Acknowledgment

The authors are grateful to the Mathematical Biosciences Institute for support. A Royal Society Wolfson Research Merit Award, EPSRC Grant EP/K022547/1 and NSF grant 0959427 are also acknowledged.

References

Baíllo, A., Cuevas, A., and Fraiman, R. (2011). Classification methods for functional data. In Ferraty, F. and Romain, Y., editors, *The Oxford Handbook of Functional Data Analysis*. Oxford University Press, Oxford. MR2917982

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.

Cheng, W., Dryden, I. L., Hitchcock, D. B., and 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.

Jones, M. C. and Henderson, D. A. (2007). Kernel-type density estimation on the unit interval. *Biometrika*, 94(4):977–984. MR2416803

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

Wu, W., Hatsopoulos, N. G., and Srivastava, A. (2014). Introduction to neural spike train data for phase-amplitude analysis. *Electron. J. Statist.*, 8:1759–1768, Special Section on Statistics of Time Warpings and Phase Variations.