MOX–Report No. 46/2013

Statistics of Time Warpings and Phase Variations

Marron, J.S.; Ramsay, J.O.; Sangalli, L.M.; Srivastava, A.

MOX, Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox@mate.polimi.it http://mox.polimi.it

Statistics of Time Warpings and Phase Variations

J.S. Marron[1], James O. Ramsay[2], Laura M. Sangalli[3] and Anuj Srivastava[4]

[1] Department of Statistics and Operations Research, University of North Carolina, NC, USA, marron@unc.edu
[2] Department of Psychology, McGill University, Montreal, Quebec, Canada, ramsay@psych.mcgill.ca
[3] MOX - Department of Mathematics, Politecnico di Milano, Milano, Italy, laura.sangalli@polimi.it
[4] Department of Statistics Florida State University, FL, USA, anuj@stat.fsu.edu

Keywords: Functional Data Analysis, Phase and Amplitude variation, Registration, Warping

Abstract

Many methods exist for one dimensional curve registration, and how methods compare has not been made clear in the literature. This special section is a summary of a detailed comparison of a number of major methods, done during a recent workshop. The basis of the comparison was simultaneous analysis of a set of four real data sets, which engendered a high level of informative discussion. Most research groups in this area were represented, and many insights were gained, which are discussed here. The format of this special section is four papers introducing the data, each accompanied by a number of analyses by different groups, plus a discussion summary of the lessons learned.

1 Introduction

Functional Data Analysis (FDA) is a popular statistical area that is maturing in both practice and theory. An important challenge to the analysis of a sample of functions or curves is to separate *amplitude variation* from *phase variation*. Amplitude variation is sometimes called "vertical variation", and phase variation is then referred to as "horizontal variation," or in the case of time, "tempo." These concepts are illustrated using a simulated example in Figure 1. The left panel shows a sample of curves, each having two peaks, with both types of variation, as seen from the peak locations and heights. The amplitude variation is shown in the center panel, which shows the same peaks after an alignment or registration process. The right panel displays

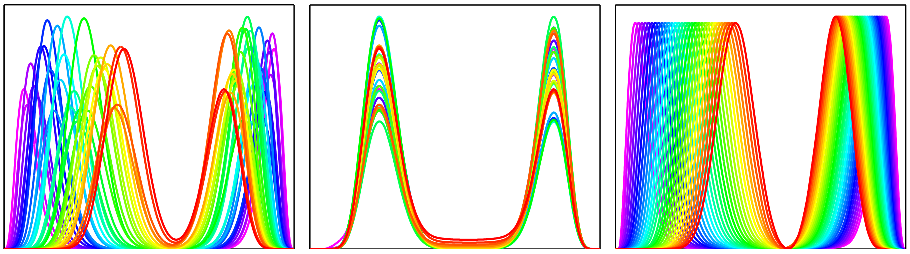

Figure 1: A simulated example, based on a sample of curves (left panel), decomposed into amplitude variation (center panel) and phase variation (right panel).

the phase variation in this example in terms of a mean curve plotted against warps or strictly monotonic transformations of the horizontal axis.

While most approaches to FDA ignore this amplitude/phase decomposition, it has become clear in range of real data analysis contexts, including growth curves, motion tracking data, chemical spectra, anatomical data and neuroscience data, that such ignorance can entail very substantial loss in statistical efficiency and interpretation. This realization has motivated a number of efforts to extract separate phase and amplitude modes of variation.

Note that there are often quite different statistical contexts where such registration of curves is useful. In some phase variation is a nuisance. This happens, for example, chemical spectral data, where peak locations representing given substances need to be aligned; but it is the amounts of substances, reflected by the heights of the peaks, that is the main focus of the analysis. In other situations, both amplitude and phase modes of variation are substantively interesting, and should be studied in terms of each type of variation separately and also in terms of joint variation. Human growth and movement data are generally of this type.

Each of the previous efforts at decomposing amplitude and phase variation typically involves the analysis of a challenging real data set where curve registration is important, and good results from use of the proposed method on that data set are shown. While that format is typical for publication in the statistical literature, it fails to provide useful comparison of methods, in particular not elucidating their relative strengths and weaknesses. This need for comparison could be met by requiring authors to do comparisons themselves, but this is generally a difficult request due to lack of availability of general purpose software implementations of existing methods and to the need for expert-level tuning that is often needed to get a good result.

In response to the need for a more global and useful comparison of these methods, the Mathematical Biosciences Institute at the University of Ohio hosted a workshop in 2012 with representation from most of the principal research groups in this area of curve registration. The initial basis of the workshop was for each group to do some analysis on a common set of four real data sets. The results from the various groups then formed the basis extensive discussion, which led in turn to many new insights. This special section aims to convey the important lessons learned at this workshop to the larger statistical community.

Following the data–centric orientation of the workshop, this special section is oriented around these four data sets:

- Proteomic data: spectral data collected for the study of Acute Myeloid Leukemia by the Adelaide Proteomics Center.
- Juggling data: records of forefinger motion during a juggling exercise, recorded from infra-red emitting diodes at McGill University.
- Spike train data: records of the electrical activity of a movement-encoded neuron in the primary motor cortex, collected in the Hatsopoulos Lab at the University of Chicago.
- AneuRisk65 data: sets of three-dimensional vascular geometries, obtained from 3D angiographies collected within the AneuRisk project for the study of cerebral aneurysms pathogenesis.

All data sets are available at the MBI website:
http://mbi.osu.edu/2012/stwdescription.html.

The main ideas behind each of these data sets, as well as data analytic goals of interest and the preprocessing that was required, is presented in the four main papers in this special section. The analyses presented by the various research groups are included in a format similar to discussions in other contexts. Next, additional relevant comments together with a summary of insights gained, are given by the original providers of the data.

2 Review

This section contains a brief discussion of a few highlights of the curve registration literature. This is not intended to be comprehensive, but instead to eliminate the need for (perhaps too repetitive) re-discussion of these papers in several of the contributions to this special section.

The classic paper is Sakoe and Chiba (1978), who developed the dynamic time warping registration algorithm for registering sequences of phonemes to a template phrase as an aid to automatic speech recognition. Time in this case was discrete, and the time points of the sequence to be aligned were transformed by a monotonic step function. These warping functions resemble in both shape and algorithm the isotonic regression functions developed at about the same time by Barlow et al. (1972).

The fact that these transformations of discrete time do not exclude either horizontal or vertical jumps renders their use problematical in functional data contexts where curves are required to be differentiable. There is now a large statistical literature on families of smooth strictly monotone curves, including basis function expansions of their log-derivatives proposed in Ramsay (1996) and used for registration in Ramsay and Silverman (2005).

Landmark registration, where a few anchor points which correspond across the family of curves are aligned, employs a monotone smoothing of the candidate curve's landmark times plotted against the corresponding template curve times. See Gasser and Kneip (1995) and Ramsay and Silverman (2005) for recent results of this type, and for access to the earlier literature.

While landmarks are useful when they exist in a natural way, in many situations these cannot be found or unambiguously located. Hence, various landmark–free approaches, which treat curves as continuous data objects, have been subject of more recent studies. Important recent results (containing many earlier references) include: Ramsay and Li (1998); Wang and Gasser (1999); Gervini and Gasser (2004); Ramsay and Silverman (2005); Kaziska and Srivastava (2007); Sangalli et al. (2009), Kneip et al. (2000); Liu and Müller (2004); James (2007).

Registration can be performed jointly with modelling and analysis of data, as in the registration to principal components method described in Kneip and Ramsay (2008).

The issue of registration can also be combined with the one of clustering functional data. Some works considering this aspect are Sangalli et al. (2010); Tang and Müller (2009); Liu and Yang (2009); Boudaoud et al. (2010).

Research providing insightful theoretical frameworks can be found in Srivastava et al. (2011b); Vantini (2012).

Important related work can also be found in the context of longitudinal data, where semiparametric non-linear mixed-effects models are proven to be useful Lawton et al. (1972); Lindstrom and Bates (1990); Ke and Wang (2001); Altman and Villarreal (2004); Brumback and Lindstrom (2004).

The issue of registration has been considered also in the shape analysis field. In particular, the earliest work on elastic shape analysis of planar curves is by Younes Younes (1999) who introduced an elastic metric and a complex square-root representation for enabling Euclidean analysis. This was followed by more elaborate studies of such representations, including Younes et al. (2008); Michor and Mumford (2006); Mio et al. (2007) and Srivastava et al. (2011a). The last paper extended this elastic shape analysis from planar curves to curves in arbitrary Euclidean spaces.

Finally, although the workshop focus has been on the registration of curves, possibly multidimensional, it is important to cite the work on registration of surfaces, in imaging. For refer the interested reader to the book of Modersitzki (2003), and references therein.

Acknowledgements

All participants are very grateful to the Mathematical Biosciences Institute (http://mbi.osu.edu/) for hosting many aspects of this effort, including data accessibility, travel funding, and an ideal venue for the personal interaction that has been the key to success.

References

Altman, N. and Villarreal, J. (2004), "Self-modelling regression for longitudinal data with time-invariant covariates," *Canad. J. Statist.*, 32, 251–268.

Barlow, R., Bartholomew, J., Brenner, J., and Brunk, H. (1972), *Statistical Inference Under Order Restriction*, Wiley, London.

Boudaoud, S., Rix, H., and Meste, O. (2010), "Core Shape modelling of a set of curves," *Comput. Statist. Data Anal.*, 54, 308–325.

Brumback, L. and Lindstrom, M. (2004), "Self modeling with flexible, random time transformations," *Biometrics*, 60, 461–470.

Gasser, T. and Kneip, A. (1995), "Searching for structure in curve samples," *Journal of the American Statistical Association*, 90, 1179–1188.

Gervini, D. and Gasser, T. (2004), "Self-modelling warping functions," *Journal of the Royal Statistical Society, Series B*, 66, 959–971.

James, G. M. (2007), "Curve alignment by moments," *The Annals of Applied Statistics*, 1, 480–501.

Kaziska, D. and Srivastava, A. (2007), "Gait-Based Human Recognition by Classification of Cyclostationary Processes on Nonlinear Shape Manifolds," *Journal of the American Statistical Association*, 102, 1114–1128.

Ke, C. and Wang, Y. (2001), "Semiparametric nonlinear mixed-effects models and their applications," *J. Amer. Statist. Assoc.*, 96, 1272–1298.

Kneip, A., Li, X., MacGibbon, K., and Ramsay, J. (2000), "Curve registration by local regression," *Can J Stat*, 28, 19–29.

Kneip, A. and Ramsay, J. O. (2008), "Combining registration and fitting for functional models," *Journal of the American Statistical Association*, 103, 1155–1165.

Lawton, W. H., Sylvestre, E. A., and Maggio, M. S. (1972), "Self Modeling Nonlinear Regression," *Technometrics*, 14, 513–532.

Lindstrom, M. J. and Bates, D. M. (1990), "Nonlinear mixed effects models for repeated measures data," *Biometrics*, 46, 673–687.

Liu, X. and Müller, H.-G. (2004), "Functional convex averaging and synchronization for time-warped random curves," *J Am Stat Assoc*, 99, 687–699.

Liu, X. and Yang, M. (2009), "Simultaneous curve registration and clustering for functional data," *Comput. Statist. Data Anal*, 53, 1361–1376.

Michor, P. W. and Mumford, D. (2006), "Riemannian geometries on spaces of plane curves," *J. Eur. Math. Soc.*, 8, 1–48.

Mio, W., Srivastava, A., and Joshi, S. H. (2007), "On Shape of Plane Elastic Curves," *Intl. Journal of Computer Vision*, 73, 307–324.

Modersitzki, J. (2003), *Numerical Methods for Image Registration*, Oxford University Press.

Ramsay, J. (1996), "Estimating smooth monotone functions," *Journal of the Royal Statistical Society, Series B.*, 60, 365–375.

Ramsay, J. O. and Li, X. (1998), "Curve registration," *Journal of the Royal Statistical Society, Series B.*, 60, 351–363.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer, 2nd ed.

Sakoe, H. and Chiba, S. (1978), "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26, 43–49.

Sangalli, L. M., Secchi, P., Vantini, S., and A.Veneziani (2009), "A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery," *J. Amer. Statist. Assoc.*, 104, 37–48.

Sangalli, L. M., Secchi, P., Vantini, S., and Vitelli, V. (2010), "K-mean alignment for curve clustering," *Computational Statistics and Data Analysis*, 54, 1219–1233.

Srivastava, A., Klassen, E., Joshi, S., and Jermyn, I. (2011a), "Shape Analysis of Elastic Curves in Euclidean Spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33, 1415–1428.

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

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

Vantini, S. (2012), "On the definition of phase and amplitude variability in functional data analysis," *TEST*, 21, 676–696.

Wang, K. and Gasser, T. (1999), "Synchronizing sample curves nonparametrically," *The Annals of Statistics*, 27, 439–460.

Younes, L. (1999), "Optimal matching between shapes via elastic deformations," *Journal of Image and Vision Computing*, 17, 381–389.

Younes, L., Michor, P. W., Shah, J., Mumford, D., and Lincei, R. (2008), "A Metric on Shape Space with Explicit Geodesics," *Matematica E Applicazioni*, 19, 25–57.

MOX Technical Reports, last issues

Dipartimento di Matematica "F. Brioschi",
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

46/2013 Marron, J.S.; Ramsay, J.O.; Sangalli, L.M.; Srivastava, A.
Statistics of Time Warpings and Phase Variations

45/2013 Sangalli, L.M.; Secchi, P.; Vantini, S.
Analysis of AneuRisk65 data: K-mean Alignment

44/2013 Sangalli, L.M.; Secchi, P.; Vantini, S.
AneuRisk65: a dataset of three-dimensional cerebral vascular geometries

43/2013 Patriarca, M.; Sangalli, L.M.; Secchi, P.; Vantini, S.
Analysis of Spike Train Data: an Application of K-mean Alignment

42/2013 Bernardi, M.; Sangalli, L.M.; Secchi, P.; Vantini, S.
Analysis of Juggling Data: an Application of K-mean Alignment

41/2013 Bernardi, M.; Sangalli, L.M.; Secchi, P.; Vantini, S.
Analysis of Proteomics data: Block K-mean Alignment

40/2013 Pacciarini P.; Rozza G.
Stabilized reduced basis method for parametrized advection-diffusion PDEs

39/2013 Ieva, F.; Marra, G.; Paganoni, A.M.; Radice, R.
A semiparametric bivariate probit model for joint modeling of outcomes in STEMI patients

38/2013 Bonizzoni, F.; Nobile, F.
Perturbation analysis for the Darcy problem with log-normal permeability

37/2013 Lever, V.; Porta, G.; Tamellini, L.; Riva, M.
Characterization of basin-scale systems under mechanical and geochemical compaction