MATH NEWSLETTER
UNIVERSITY OF COPENHAGEN

HELLE SØRENSEN
AUGUST 2016

My research: Functional data analysis

Functional data analysis (FDA) is the area of statistics which is concerned with data consisting of curves, surfaces or other objects varying over a continuum (typically time or location). Work on statistical methods for functional data was already initiated by the 60's but development accelerated in the 90's as functional data became more widely produced and recognized. The ever increasing complexity of data continue to provide new challenges for data analysis, and FDA is a lively research area these days.

In practice, data from an individual (subject i, say) consist of discrete measurements $x_{i1},\dots,x_{iN}$ taken at time or location points $s_{i1},\dots,s_{iN}$, but these data points are assumed to arise from a (smooth) function X_i such that x_{ij} is an observation of $X_i(s_j)$, and the interest is in the functions as such rather than the individual measurements. In order to do statistics we thus need probability distributions on suitable function spaces (although finite-dimensional basis systems are often used to represent the functions).

A typical dataset consists of functions from several individuals together with other measurements and information, and the aims of statistical analyses of functional data are similar to those with other types of data. For example: *What is the association, if any, between functional variables and other variables in the dataset?* (a regression problem), *For a new functional observation, which group does it come from?* (a classification problem), *Do functions from different groups of individuals differ in a systematic manner, and on which subintervals?* (comparison of groups), etc. In any case, the challenge is to take into account the functional nature of the data in appropriate ways.

In the following I will describe a recent project on classification. This is joint work with Seyed Nourollah Mousavi who graduated as PhD from UCPH 2016.

Multinomial functional regression

Consider the problem of classification with a functional predictor. Data consists of functions from n individuals, $x_i : (0,1) \to \mathbb{R}$, $i = 1,\dots,n$. The n individuals come from M different groups, and the group membership is known for the n individuals. The aim is to identify the group membership for a *new curve*.

We need to establish a model for the relationship between groups and curves, and use a *functional multinomial regression model* for that purpose. For a given function $x : (0,1) \to \mathbb{R}$, let $p_m(x)$ be the conditional probability that the function x belongs to group m, and assume that it takes the form

$$p_m(x) = \frac{e^{\alpha_m + \int \beta_m(t)x(t)dt}}{\sum_{l=1}^{M} e^{\alpha_l + \int \beta_l(t)x(t)dt}}.$$

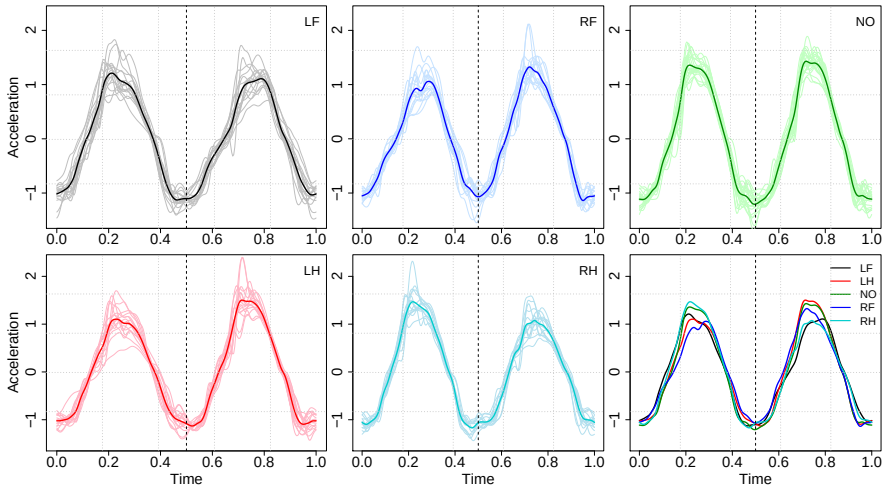
where $\alpha_1,\dots,\alpha_M$ are unknown constants and $\beta_1,\dots,\beta_M : (0,1) \to \mathbb{R}$ are unknown functions which must be estimated from the data. All integrals are over the unit interval.

We use an estimation approach based on wavelets. Each covariate function x_i is expanded in a wavelet basis, and the wavelet coefficients are used as covariates in a penalized multinomial regression with L^1 penalization (LASSO). The penalization implies that only a subset of the wavelet coefficients are used actively in the regression model. In the end, the estimated coefficients are translated to estimated intercepts $\hat{\alpha}_1,\dots,\hat{\alpha}_M$ and estimated coefficient functions $\hat{\beta}_1,\dots,\hat{\beta}_M$.

Once the α's and β's have been estimated, it is easy to use the corresponding regression model for prediction: For a new function $\tilde{x}$, first compute the linear predictors $\hat{\alpha}_m + \int \hat{\beta}_m(t)\tilde{x}(t)dt$ and then the corresponding conditional probabilities $\hat{p}_1(\tilde{x}), \ldots, \hat{p}_M(\tilde{x})$, and finally assign $\tilde{x}$ to the group with the largest probability.

There is always a risk of „overfitting“, that is, a risk that the model finds spurious relationships that are not representative for new data. In order to evaluate the quality of classification procedures, data are therefore often split into *training data* and *test data*. The training data are used to estimate the α's and β's. The corresponding model is then used to classify the test data, and the proposed groups and the true groups are compared. For small datasets a *leave-one-out* evaluation is sometimes carried out: For observation i the dataset consisting of all curves except curve i is used as training data, and curve i is used as test data. This is repeated for all $i = 1, \ldots, n$.

For an application, consider the figure below. It shows a total of 85 acceleration signals collected from horses while trotting. There are five groups corresponding to lameness on each of the four limbs (LF = left fore, LH = left hind, RF = right fore, RH = right hind), and no lameness (NO = normal). Thin lines show the 85 signals whereas thick lines show the average curves in each group (also shown in the bottom right panel). The data were collected by Maj Halling Thomsen from Department of Large Animal Sciences, UCPH.



In this application classification of a new curve is equivalent to diagnosis: *Is the horse lame and, if so, on which limb?* A leave one-one-out analysis gave the results listed in the table below:

	Predicted group				
True group	LF	LH	NO	RF	RH
LF	13	0	2	0	1
LH	1	13	1	1	0
NO	0	1	20	1	1
RF	0	1	1	14	0
RH	1	0	2	0	11

The 71 signals in the diagonal correspond to correctly classified signal, yielding a success rate of 84%. We compared to several classification procedures from the literature, and our method gave at least as good results both for the horse data and for another dataset on speech recognition.