Workshop: High dimensional and dependent functional data

Hans-Georg Müller, University of California, Davis

Modeling repeatedly observed functional data

Abstract: Repeatedly observed and thus dependent functional data are encountered when random curves are recorded repeatedly for each subject in a sample. The proposed models lead to a flexible yet interpretable and straightforward decomposition of the inherent variation in repeatedly observed functional data and are implemented through a two-step functional principal component analysis. The time points where functions are recorded may be irregular and sparse as is often the case in longitudinal studies. The estimated model components are shown to be consistent in various scenarios. The methods are illustrated through the analysis of longitudinal mortality data from period lifetables and also through simulation studies.

This talk is based on joint work with Kehui Chen, University of Pittsburgh.

Fabian Scheipl, Ludwig-Maximilians-Universität München

Additive mixed models for correlated functional data

Abstract: We propose a comprehensive framework for flexible additive regression models for correlated functional responses, allowing for multiple partially nested or crossed functional random effects with spatial, temporal, or longitudinal correlation structures. Additionally, our framework includes linear effects of functional covariates and linear or smooth effects of scalar covariates that can vary smoothly over the index of the functional response and accommodates densely or sparsely observed functional responses and predictors which may be observed with additional error. Inference in this framework can be performed by standard software for generalized additive models (GAMs), allowing us to take advantage of established robust and flexible estimation algorithms. Simulation experiments show that the proposed method recovers relevant effects reliably, handles small group sizes and/or low numbers of replications well and also scales to larger data sets. Application examples on spatial and longitudinal functional data demonstrate that the proposal yields flexible model specifications that do justice to complex data situations and yield interpretable results.

Jan Gertheiss, Ludwig-Maximilians-Universität München

Longitudinal scalar-on-functions regression with application to tractography data

Abstract: We propose a class of estimation techniques for scalar-on-function regression in longitudinal studies where both outcomes and functional predictors may be observed at multiple visits. Our methods are motivated by a longitudinal brain diffusion tensor imaging (DTI) tractography study. One of the primary goals of the study is to

evaluate the contemporaneous association between human function and brain imaging over time. The complexity of the study requires development of methods that can simultaneously incorporate: (1) multiple functional (and scalar) regressors; (2) longitudinal outcome and functional predictors measurements per patient; (3) Gaussian or non-Gaussian outcomes; and, (4) missing values within functional predictors. We propose two versions of a new method, longitudinal functional principal components regression. These methods extend the well-known functional principal component regression and allow for different effects of subject-specific trends in curves and of visit-specific deviations from that trend. The different methods are compared in simulation studies, and the most promising approaches are used for analyzing the tractography data.

John Moriarty, University of Manchester

Gaussian process regression on a phylogeny

Abstract: Biological data objects often have both of the following features: (i) they are functions rather than single numbers or vectors, and (ii) they are correlated due to phylogenetic relationships. I will give a flexible statistical model for such data, by combining assumptions from phylogenetics with Gaussian processes. The model provides a prior distribution for Bayesian inference, enabling both prediction (for ancestral functions) and model selection. This work extends the popular phylogenetic Brownian Motion and Ornstein-Uhlenbeck models to functional data, and extends Gaussian Process regression to phylogenies. Joint work with Nick Jones (Imperial College).

Alessandra Menafaglio, MOX - Politecnico di Milano

Universal kriging prediction for spatially dependent functional data of a Hilbert space

Abstract: The problem of analyzing and predicting spatially dependent functional data is addressed proposing an extension of some geostatistical tools to nonstationary functional random fields, with a Functional Data Analysis approach. Having defined proper measures of global spatial variability for functional random processes, an extension of the Universal Kriging method to functional data belonging to a Hilbert space is proposed.

Consistently with the new established theoretical results, a three-step procedure for the prediction of non-stationary spatial dependent functional data is proposed: model selection for spatial mean (drift), decomposition of the observed process into a drift term (assumed deterministic) and a residual stochastic process, Universal Kriging prediction.

A simulation study is performed in order to test the behavior of the developed methodology in different scenarios. The proposed procedure is finally applied to daily mean temperature curves recorded in 35 meteorological stations located in Canada's Maritime Provinces.

Łukasz Kidziński, Université Libre de Bruxelles

Estimation in Hilbertian linear models

Abstract: Consider estimation of the operator Ψ in the linear model $Y_k = \Psi(X_k) + \varepsilon_k$, when X and Y take values in Hilbert spaces H_1 and H_2, respectively. Our main objective is to obtain a consistent estimator of Ψ under as mild as possible assumptions.

The crucial difficulty in this problem is that we are working with an infinite dimensional operator Ψ, which needs to

be approximated by a sample version $\hat{\Psi}_K$ of finite dimension K. A delicate issue is then the choice of K. In existing papers determination of K requires very specific assumptions on the spectrum of the covariance operator of the regressor variables. We explain why such assumptions pop up throughout the literature and show that consistency can be established without any assumption on the spectrum by proposing a purely data-driven procedure for the choice of K.

We obtain our results in the framework proposed by S. Hormann and P. Kokoszka [1]. In particular, we allow the regressors X_k to be weakly dependent which for example allows us to include the case of a functional autoregressive model. This model has been intensively investigated in the functional literature and is often used to model autoregressive dynamics of a functional time series (see e.g. Bosq [2]). We can not only greatly simplify the assumptions needed for consistent estimation but also allow for a more general setup.

Since all the quantities involved can be computed from the sample our procedure can be a fast alternative to cross-validation or AIC type criteria for the choice of K in practical applications. We compare the performance of the proposed estimators in a small simulation study.

[1] S. Hörmann and P. Kokoszka, Functional Time Series, Handbook of Statistics, forthcoming.

[2] D. Bosq, Linear Processes in Function Spaces, Springer, New York, 2000.

Denis Bosq, LSTA, Université Pierre et Marie Curie - Paris VI

Constructing functional linear filters

Abstract: This talk deals with prediction in large dimensions. The goal is to construct the best linear predictor of Y given X, say $\lambda(X)$, where X and Y are random variables with values in two Hilbert spaces. The difficulty comes from the fact that, in general, λ is not continuous, since it is not defined everywhere.

In a first part we study general forms of ARMA processes in a Hilbert space, associated with noncontinuous linear operators.

In the second, by using measurable linear transformations and linearly closed subspaces we obtain explicit forms of the best linear predictor.

Various examples are considered: processes with roots of modulus 1, Kalman-Bucy filter, compound Ornstein-Uhlenbeck process, model with noise, tensorial product of Gaussian random variables, extended exponential smoothing, Bayesian estimation in Hilbert spaces.

Jairo Cugliari, INRIA

Conditional autoregressive Hilbertian process

Abstract: We focus on the problem of predicting a function-valued stochastic process. The approach we adopt is based on the notion of Autoregressive Hilbert (ARH) processes. Estimation and prediction of arh processes impose interesting challenges due to the infinite dimension of the space where the process is defined.

If additional exogenous information is available, we may want to use it in for estimation and prediction purposes.

We aim here at introducing an exogenous covariate in the arh process in such a way that conditionally on the covariate the process becomes an arh. We call the new process Conditional Autoregressive Hilbertian process (carh).

In addition to its definition, estimation and prediction procedures are proposed. We give consistency results that justifies the theoretical relevance of our propositions. Finally, we perform numerical experiments on simulated data as well as on real data.

Maurice Berk, Imperial College London

Multi-level functional principal components analysis models for replicated genomics time series data sets

Abstract: FDA methods have proven to be extremely useful in modelling replicated genomics time series data sets, resulting in genuinely novel biological insight. In order to minimise both the computational burden and the complexity of the models, standard practice with such data sets is to ignore the between gene-variation and to fit each gene independently with a functional mixed-effects model that accounts only for between-replicate variation. In this talk I will present work to investigate the impact such an assumption has on our ability to model the data well, and suggest some reasons why no one has attempted to address it. I will then introduce a novel skew-t-normal multi-level functional principal components analysis model that can be used to simultaneously estimate both between-gene and between-replicate variation before giving some thoughts on the practical usefulness of this model and the prospect for future developments.

Victor Panaretos, Ecole Polytechnique Fédérale de Lausanne (EPFL)

Fourier analysis of stationary functional data

Abstract: We consider the problem of drawing statistical inferences on the second-order structure of weakly dependent functional time series. Much of the research in functional time series has focused on inference for stationary time series that are linear. In this talk we consider the problem of inferring the complete second-order structure of stationary functional time series without any structural modeling assumptions. Our approach is to formulate a frequency domain framework for weakly dependent functional data, employing suitable generalisations of finite-dimensional notions. We introduce the basic ingredients of such a framework, propose estimators, and study their asymptotics under functional cumulant-type mixing conditions. Based on joint work with Shahin Tavakoli (EPFL).

Valentin Patilea, CREST-Ensai

Projection-based nonparametric goodness-of-fit testing of regression models with scalar and functional covariates

Abstract: This paper studies the problem of nonparametric checks of regression models with scalar responses and hybrid covariates, that is both scalar and functional covariates. More precisely we study (a) the significance of a functional covariate in a parametric regression model with a finite-dimension vector of covariates; and (b) the goodness-of-fit of the partial functional linear regression model. The functional covariate takes values in $L^2[0,1]$, the Hilbert space of the square-integrable real-valued functions on the unit interval. Our test is based on the remark that checking the no-effect of a functional covariate on a scalar random variable is equivalent to checking the nullity of the conditional expectation of the error term given a sufficiently rich set of projections of the covariate. Such projections could be on elements of norm 1 from finite-dimension subspaces of $L^2[0,1]$. Next, the idea is to search a

finite-dimension element of norm 1 that is, in some sense, the least favorable for the null hypothesis. Finally, it remains to perform a nonparametric check of the nullity of a conditional expectation given the finite dimension covariates and the scalar product between the covariate and the selected least favorable direction. For such finite-dimension search and nonparametric check we use a kernel-based approach. As a result, our test statistic is a quadratic form based on kernel smoothing and the asymptotic critical values are given by the standard normal law. The test is able to detect nonparametric alternatives. The error term could present heteroscedasticity of unknown form. We do no require the law of the covariate X to be known. The test could be implemented quite easily and performs well in simulations and real data applications. Extension to the case of dependent data is also discussed. Results related to those proposed in the talk could be found in the paper arXiv:1205.5578v1 [math.ST].

The proposed results are based on joint work with Cesar Sanchez-Sellero.

Davide Pigoli, MOX - Department of Mathematics, Politecnico di Milano

Distances and inference for covariance functions

Abstract: Data is increasingly becoming available that is best described as being functional. A framework will be presented for providing inference concerning the covariance operator of a functional random process, where the covariance operator itself is an object of interest for the statistical analysis. While some finite dimensional distances for comparing positive definite covariance matrices naturally lend themselves to functional analogues, others do not have natural extensions. Having considered possible distance for covariance functions, some distance-based inferential techniques are proposed.

First, a Frechét estimator for the average covariance function is introduced. Then, a permutation procedure to test the equality of the covariance operator between two groups is considered.

Finally, the proposed techniques will be applied to a real problem, concerning relationships among Romance languages. In the linguistic analysis of human speech, the overall mean structure of the data produced is often not of interest, but rather the variations that can be found within the language. Here it will be shown that different languages can be compared and even predicted through the proposed distances between covariance functions, allowing, for the first time, a quantitative analysis of language relations based on speech recordings rather than discrete textual analysis.

[Joint work with John Aston, Ian Dryden and Piercesare Secchi]

Nicole Augustin, University of Bath

Modelling fat mass as a function of weekly physical activity profiles measured by Actigraph accelerometers

Abstract: In the epidemiological setting where physical activity may be a health outcome or the predictor of a health outcome, high dimensional accelerometer activity counts are summarised into a single summary statistic per individual, e.g. total activity, defined as the average accelerometer counts per minute, average daily moderate to vigorous physical activity (MVPA), defined as the average minutes per day spent at moderate or vigorous activity, or average sedentary behaviour, defined as the average minutes per day spent in sedentary activity (Riddoch et al., 2009; Mitchell et al., 2009, 2011). For each of these summaries cut-points of counts per minute are used corresponding to light, moderate and vigorous activity, either based on findings in the literature or on calibration studies performed

with a subset of the study population (Mattocks et al., 2007). Using only scalar summaries ignores the pattern of physical activity, meaning the distribution of the activity counts and the patterns of activity counts through time. That is, a large amount of information is lost and this is a waste of resources.

We show results on the Avon longitudinal study of parents and children (ALSPAC) using a new approach for modelling the relationship between health outcomes and physical activity measured by accelerometers. The key feature of the model is that it utilises the full profile of measured physical activity, rather than traditionally used scalar summary measures. In order to compare accelerometer profiles between individuals and to reduce the high dimension of the profiles, a functional summary of the profiles is used. We consider the histogram as a functional summary due to its simplicity and ease of interpretation. The model used is a generalised regression of scalars on functions. Our results indicate that the effect of physical activity is not constant over the activity range.

[Joint work with Calum Mattocks, Ashley R. Cooper, Andy R. Ness and Julian J. Faraway]

References

Mattocks C, Leary S, Ness A, Deere K, Saunders J, Tilling K, Kirkby J, Blair S and Riddoch C 2007 International Journal of Pediatric Obesity pp. 1–9.

Mitchell J, Mattocks C, Ness A, Leary S, Pate R, Dowda M, Blair S and Riddoch C 2009 Obesity (Silver Spring) 17(8), 1596–1602.

Mitchell J, Pate R, Dowda M, Mattocks C, Riddoch C, Ness A and Blair S 2011 Medicine and Science in Sports and Exercise.

Riddoch C, Leary S D, Ness A R, Blair S N, Deere K, Mattocks C, Griffiths A, Smith G D and Tilling K 2009 British Medical Journal 339, :b4544.

Elizabeth Sweeney, Johns Hopkins Bloomberg School of Public Health

Automatic lesion incidence estimation and detection in multiple sclerosis using multisequence longitudinal magnetic resonance images

Abstract: Detecting incidence and enlargement of lesions is essential in monitoring the progression of multiple sclerosis (MS). In clinical trials, lesion load is observed by manually segmenting and comparing serial magnetic resonance images (MRIs), which is time consuming, costly, and prone to inter- and intra- observer variability. Subtracting images from consecutive time points nulls stable lesions, leaving only new lesion activity. We propose Subtraction-Based Logistic Inference for Modeling and Estimation (SuBLIME), an automated method for segmenting incident lesion voxels.

We use logistic regression models incorporating multiple MRI sequences and subtraction images from consecutive longitudinal studies to estimate voxel-level probabilities of lesion incidence. We use T1-weighted, T2-weighted, fluid attenuated inversion recovery (FLAIR) and proton density (PD) volumes from a total of 110 MRI studies from 10 subjects.

To assess the performance of the model, we assign five subjects to a training set and the remaining five to a validation set. With SuBLIME, lesion incidence is detected and delineated in the validation set with an area under the receiver operator characteristic curve of 99voxel level.

This fully automated and computationally fast method allows sensitive and specific detection of lesion incidence that

can be applied to large collections of images. Using the explicit form of the statistical model, SuBLIME can easily be adapted to cases when more or fewer imaging sequences are available.

Haipeng Shen, University of North Carolina at Chapel Hill

Nonparametric independent component analysis for colored sources with mixed spectra

Abstract: Independent component analysis (ICA) has been a powerful data-driven method for blind source separation. In applications such as functional neuroimaging analysis, the independent components (ICs), or sources, exhibit certain correlation structures, i.e. are colored. We develop a frequency-domain ICA method within the framework of Whittle likelihood that allow the colored sources to have possibly mixed spectra, i.e. a mixture of line spectra and spectral density functions. We model the mixed spectra using a combination of indicator functions and cubic splines. Parameters for the mixed spectra and the corresponding mixing matrix are then estimated via maximum Whittle likelihood. The performance of the method is demonstrated numerically through real applications and simulation studies.

Ci-Ren Jiang, Academia Sinica

Nonparametric response function estimation via FPCA with an application to Dynamic PET Data

Abstract: In dynamic PET data analysis, injected radioactive tracer concentrations are measured over time to help understand functional processes in the body. Traditionally, parametric forms are assumed for the implied impulse response functions while estimating the concentration; however, these parametric assumptions are very difficult to verify and may well not hold. Therefore, we propose a nonparametric approach to estimate the response functions and thus the concentration. First, we employ FPCA with a multiplicative structure to represent the signal function for each voxel using a Karhunen-Loeve decomposition. As convolution can be viewed as a linear operator, we secondly apply deconvolution to the mean and eigenfunctions of the voxel signals. Then, the response function for each voxel can be represented as a linear combination of deconvolved mean function and deconvolved eigenfunctions where the linear coefficients are identical to the multiplicative coefficients and principal component scores in the first step. Therefore, the integral of the concentration in a finite time interval (a quantity of particular interest) can be obtained easily. This approach is demonstrated with simulation studies and real data analysis.

Laura Sangalli, MOX - Politecnico di Milano

Spatial regression models with differential regularization

Abstract:

Interfacing statistical methodology and numerical analysis techniques, we propose regression models with a partial differential regularization, that accurately estimate surfaces and spatial fields. In particular, the proposed models are able to deal with data scattered over complex bi-dimensional domains, including domains with irregular shapes and holes; they allow for spatially distributed covariate information and can impose various conditions over the boundaries of the domain. Accurate surface estimation is achieved resorting to finite elements and full uncertainty quantification is provided via inferential tools. Important extensions of the proposed models include the possibility to deal with data distributed over non-planar domains and to incorporate a priori information about the spatial structure of the

phenomenon under study.

Laura Azzimonti, MOX - Politecnico di Milano

PDE regularized blood velocity estimation

Abstract: We propose a novel functional data analysis technique for surface estimation over irregularly shaped regions, based on a multi-dimensional generalization of smoothing splines. The surface estimate is obtained via minimization of a penalized sum-of-square-error functional where the roughness penalty consists in the L2 norm of a second order partial differential operator. The method is especially well suited for applications where the prior knowledge of the problem suggests a partial differential operator modeling to some extent the phenomenon under study and its spatial dependence structure. The surface estimator is well defined, consistent and asymptotically normally distributed. Computations are done resorting to the Finite Element method. Classic inferential tools and uncertainty quantification for the estimate are derived due to the linearity of the estimator in the observations.

The application driving our research concerns the estimation of the blood-flow velocity field in a section of a carotid artery, using data provided by eco-color Doppler; in this application the differential operator considered is an approximation of the Navier-Stokes equations.

Bree Ettinger, MOX - Politecnico di Milano

Spatial regression models over two-dimensional Riemannian manifolds

Abstract: We propose a regression model for data spatially distributed over two-dimensional manifolds. In particular, our method addresses the case where the spatial data occur on an embedded surface in three-dimensions. Our approach consists of two phases: first we conformably map the original surface domain to a region in $\mathbb{R}^2$ then apply existing spatial regression techniques for planar domains, suitably modified to account for the domain deformation. The driving application for the proposed approach is the modeling of hemodynamic data, such as wall shear stress or blood pressure, generated by the blood flow on the wall of a carotid artery. These data are obtained from the computational fluid dynamics on real geometries of carotid arteries reconstructed from three-dimensional angiographies.

Jan Johannes, Université Catholique de Louvain

Adaptive estimation in functional linear models

Abstract: We consider the nonparametric estimation of the slope function in functional linear regression, where scalar responses are modeled in dependence of random functions. The theory in this presentation covers both the estimation of the slope function or its derivatives (global case) as well as the estimation of a linear functional of the slope function (local case). We propose an estimator of the slope function which is based on dimension reduction and additional thresholding. Moreover, replacing the unknown slope function by this estimator we obtain in the local case a plug-in estimator of the value of a linear functional evaluated at the slope. It is shown that in both the global and the local case these estimators can attain minimax optimal rates of convergence up to constant. However, the estimator of the slope function requires an optimal choice of a tuning parameter with regard to certain characteristics of the slope function and the covariance operator associated with the functional regressor. As these are unknown in practice, we investigate a fully data-driven choice of the tuning parameter which combines model selection and

Lepski's method inspired by the recent work of Goldenshluger and Lepski (2011). It is shown that the adaptive estimator with data-driven choice of the dimension parameter can attain the lower minimax risk bound in the global case up to a constant and in the local case up to a logarithmic factor, and this over a variety of classes of slope functions and covariance operators.

References:

[1] A. Goldenshluger and O. Lepski. Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. The Annals of Statistics, 39:1608-1632, 2011.

[2] H. Cardot and J. Johannes. Thresholding projection estimators in functional linear models. Journal of Multivariate Analysis, 101(2):395-408, 2010.

[3] F. Comte and J. Johannes. Adaptive functional linear regression. Technical report, arXiv:1112.2509 (under revision) 2012.

[4] J. Johannes and R. Schenk. Adaptive estimation of functionals in functional linear regression. Technical report, arXiv:1112.2855 (under revision) 2012.

Jian Qing Shi, Newcastle University, UK

Gaussian process regression analysis for large functional data

Abstract: The model based on Gaussian process (GP) prior and a covariance kernel can be used to fit nonlinear data with multi-dimensional functional covariates. It has been used as a flexible nonparametric approach for curve fitting, classification, clustering and other statistical problems, and has been widely applied to deal with complex nonlinear system in many different areas. In this talk, I will first introduce this GP regression analysis model and then focus on the problem when the model is used for the large scale data sets and high dimensional data. Specifically, a penalized likelihood framework will be applied to the model based on Gaussian processes. Different penalties, their ability in application given to suit the characteristics of GP models and the asymptotic properties will be discussed. Several applications to real bio-mechanical and bioinformatics data sets will also be presented.

Surajit Ray, Glasgow University, Boston University

Reconstructing trajectories of correlated functional data

Abstract: Functional Principal Components have been widely used to reconstruct trajectories with sparse observations. But in many practical applications these trajectories are correlated. In this paper, we extend the conditional expectation method (PACE) in Yao et al. (2005) to the case where sample curves are correlated. We call the proposed method SPACE which stands for spatial PACE. This methodology will be motivated by and applied to reconstructing vegetation index curves for spatially correlated remote sensing observations.

Frédéric Ferraty, Institute de Mathématiques de Toulouse, Université Paul Sabatier

Nonparametric variable selection and FDA

Abstract: The high dimensional setting is a modern and dynamic research area in Statistics. It covers numerous situations where the number of explanatory variables is much larger than the sample size. This particular setting corresponds to the observation of a collection of curves, surfaces, etc; this corresponds to the so-called functional

data. Last twenty years have been devoted to develop successful methodologies (mainly in the linear setting) able to manage such high dimensional data (HDD-I) by taking into account their continuous features. In parallel, when considering high-dimensional data as a small sample of large vectors derived from a large set of covariates (HDD-II), for instance microarray in genomics, sparse linear modelling involving variable selection techniques has been proposed to handle such high-dimensional statistical problem. In a first part of this talk, we extend the sparse linear modelling to the nonparametric setting by proposing two nonparametric variable selection algorithms. These sparse nonparametric regression methods are illustrated on some genomics data (HDD-II) which highlights the possible advantage one can expect in comparison with more conventional sparse linear alternatives. In the second part, we propose to study functional data (HDD-I) by using the nonparametric variable selection approach. Surprisingly, in some cases such pointwise methods may outperform standard continuous alternatives (for instance the functional nonparametric regression) while giving interpretable outputs. In this way, nonparametric variable selection may be a useful complementary tool for FDA.

Carlo Sguera, Universidad Carlo III, Madrid

Spatial depth-based classification for functional data

Abstract: Functional data are becoming increasingly available and tractable because of the last technological advances. We enlarge the number of functional depths by defining two new depth functions for curves. Both depths are based on a spatial approach: the functional spatial depth (FSD), that shows an interesting connection with the functional extension of the notion of spatial quantiles, and the kernelized functional spatial depth (KFSD), which is useful for studying functional samples that require an analysis at a local level. Afterwards, we consider supervised functional classification problems, and in particular we focus on cases in which the samples may contain outlying curves. For these situations, some robust methods based on the use of functional depths are available. By means of a simulation study, we show how FSD and KFSD perform as depth functions for these depth-based methods. The results indicate that a spatial depth-based classification approach may result helpful when the datasets are contaminated, and that in general it is stable and satisfactory if compared with a benchmark procedure such as the functional k-nearest neighbor classifier. Finally, we also illustrate our approach with a real dataset.

Posters

Pantelis Hadjipantelis, University of Warwick

Functional phylogenetic Gaussian process regression

Abstract: This work explores very high dimensional data, containing spatial and temporal structure, with the goal of inferring information about their phylogenetic properties; namely their rates of evolution as well as their ancestral states. A framework utilizing Gaussian process regression is presented in an attempt to merge Evolutionary Biology notions and Machine Learning methodologies. After illustrating how to move from higher to lower dimensional spaces using a combination of Principal Components and Independent Components analysis, we define an Ornstein-Uhlenbeck process diffusing through evolutionary time (a phylogeny). Thereafter, the inferential procedure follows from common marginalization and hyper-parameters estimation routines. Importantly, the current model not only is simple and well grounded mathematically but implements a number of biological intuitions and principals that render it biologically

relevant.

This is joint work with J. Aston (Warwick Statistics), J.Moriarty (Manchester Mathematics), C.Knight, D.Springate (Manchester Life Sciences) and D. Pigoli (Politecnico di Milano Mathematics).

Sophie Dabo, Université Charles De Gaulle, Lille 3

Classification and change points detection in functional data

Abstract: Recent advances in functional data analysis allow to construct different classification methods, based on the comparison between centrality curves or using change points detection.

We review some procedures that have been used to classify functional data. The main point is here to show the good practical behaviors of these procedures on a sample of curves. In addition, new theoretical advances on functional estimations related to change points detection methods are provided.

Michelle Carey, University of Limerick

A generalised smoother for linear ODEs

Abstract: Generalised smoothing aims to obtain an estimated functional entity that adheres to the data and incorporates domain specific information defined by an ODE. In this paper, we present an alternative representation of B-spline basis functions in terms of the underlying polynomials that comprise the B-spline. This perspective produces generalised penalties which can be written explicitly in terms of the parameters of the ODE. Thus, eliminating the need for parameter cascading (Cao and Ramsay (2007)). The joint estimation procedure developed is shown to produce estimates that have a higher accuracy and are more computationally efficient than estimates developed by existing methods.

Manuel Oviedo de la Fuente, Universidade de Santiago de Compostela

Functional response models in R

Abstract: A regression model is said to be "functional" when at least one of the involved variables (either a predictor variable or response variable) is functional. The case of functional response models (FRM) is analyzed by some authors: Ramsay and Silverman (1997) , Faraday (1997), Chiou, Müller and Wang (2004) and Ferraty, Laksaci, Tadj and Vieu (2011). This work is devoted to functional regression models where the response variable is functional and at least, there is one functional covariate. It focus on the case of a functional response prediction in practice, which remains largely unexplored.

The function linmod of fda package (Ramsay, Wickham, and Hooker (2011)) is a basic reference to fit or predict functional responses in R, where the functional data are represented in a basis expansion function restricted to the space of L2. An alternative approach models has proposed by Chiou, Muller and Wang (2003) where the functional responses are predicted via the Principal Analysis by Conditional Estimation (see PACE package-http://anson.ucdavis.edu/ ntyang/PACE/ <http://anson.ucdavis.edu/(2011) presented theoretical details for Kernel regression with functional responses. Ferraty and Vieu (2006) extend these ideas to build functional pseudo-

confidence area by nonparametric functional regression model (R code are downloadable athttp://www.math.univ-toulouse.fr/staph/npfda). This work also considers the refund package (Crainiceanu and Goldsmith (2011)) which implemented additive regression for functional and scalar covariates and functional responses.

The aim of this work is to compare the previous functional response models available with the functions of R package fda.usc which provides a broader, flexible tool for the analysis of functional data.

Haochang Shou, Johns Hopkins University

Structured functional principal components analysis

Abstract: Motivated by modern observational studies, we introduce a wide class of functional models that expands classical nested and crossed designs. Our approach targets a better interpretability of level specific features through natural inheritance of designs and explicit modeling of correlations. We define functional variance components as covariance operators of the latent processes and base our inference on functional quadratics and their relationship with underlying covariance structure. For modeling functions sampled at an ultra high frequency, we develop a computationally fast and scalable estimation procedure. We illustrate the methods with three data sets that represent a new generation of functional data: a high-frequency accelerometer data collected for estimating daily energy expenditure, pitch linguistic data used for phonetic analysis, and EEG data representing electrical brain activity during sleep.

Shahin Tavakoli, Ecole Polytechnique Fédérale de Lausanne (EPFL)

On mixing conditions for asymptotics in functional time series

Abstract: Motivated by DNA minicircle data from Molecular Dynamics, we investigate mixing conditions that enable us to draw statistical inferences for stationary functional data. We are interested in general stationary processes as opposed to linear processes. We review existing functional mixing conditions, examples of processes that satisfy them, and asymptotic results they allow for. We then consider moment-based functional mixing conditions, and show how these can be used to recover or extend existing asymptotic results. We also consider examples of functional processes satisfying our mixing conditions, and probe the stability of our conclusions under discrete observation of the functional time series. (based on joint work with Victor M. Panaretos, EPFL)

Jona Cederbaum, Ludwig-Maximilians-Universitaet, Munich

Functional linear mixed models for sparsely or irregularly sampled data

Abstract: Functional data which involves additional correlation structure pose a problem for conventional regression approaches. Recently, a functional analogue to the (scalar) linear mixed model [2] which is frequently used to analyze scalar correlated data has been proposed by [1]. In that work, functional principal component analysis (e.g. [3],[4]) is extended to serve as a computationally efficient estimation method for a decomposition of the covariance of correlated functional data.

Like the majority of methods in functional data analysis, the functional linear mixed model (FLMM) of [1] has been developed for grid data, e.g. it requires a large number of regularly spaced measurements per curve which is restrictive in practice.

Our motivating example for extensions to non-grid data comes from phonetic research. Linguists are interested in the change of articulation when certain consonants follow each other. In our application, different words are read out loud by subjects and their tongue movement is summarized in an one-dimensional functional index. A standardization of the different reading durations results in irregularly spaced measurements of the index between curves. Each of the words is read out up to 5 times by each proband, leading to correlated, repeated measurements both for each word and for each proband. Therefore, FLMMs based on crossed designs of the form

$$Y_{ijh}(t) = \eta(t) + B_i(t) + C_j(t) + U_{ijh}(t) + \epsilon_{ijh}(t),$$

present an adequate modeling tool. Here, Y denotes the index, eta contains covariates effects, such as the order in which the consonants occur, as well as a fixed mean effect. As an analogue to the random intercept in classical linear mixed models, $B_i(t)$ and $C_j(t)$ denote random functional intercepts for the words and the speakers, respectively. $U_{ijh}(t)$ is a word-, speaker-, and visit-specific random deviation and epsilon is random homoscedastic white noise.

So far, the case of irregularly observed and sparse functional data has received little attention. This may be due to the increased complexity of the implementation and computational challenges. For example, it is not possible to smooth each sparsely observed function separately as can be done in the case of grid data since too few observations may be available per curve.

In this work, we develop an estimation procedure for FLMMs which is applicable to sparsely or irregularly sampled data and therefore extends the approach of [1]. We conduct a simulation study to compare different implementations and present a case study on phonetic data. The procedure is not restricted to longitudinal functional data, but allows for more general correlation structures such as crossed designs. Our approach builds on existing methods in functional data analysis as well as on mixed model theory. It can be seen as a generalization of mixed models to functional data as well as of the FLMM to non-grid data.

References

[1] Greven, Crainiceanu, Cao, and Reich. Longitudinal functional principal component analysis. Electronic Journal of Statistics, 4, 2010.

[2] Laird and Ware. Random-effects models for longitudinal data. Biometrics, 38(4), 1982.

[3] Ramsay and Silverman. Functional data analysis. Springer series in statistics. Springer, 2005.

[4] Rice and Silverman. Estimating the Mean and Covariance Structure nonparametrically when the Data are Curves. Journal of the Royal Statistical Society, Series B 53(1), 1991.

AhYeon Park, University College London

Principal fitted components for functional regression

Abstract: We propose an extension of principal fitted components (PFCs) to functional data that allows a simultaneous smoothing step. In functional principal components regression, we project the random function X onto the subspace spanned by the first few functional principal components (FPCs), prior to regressing the scalar y on the function X. However, FPCs are entirely determined by the marginal distribution of X, and may not be related to the response y. This motivates us to generalise the existing multivariate approach based on the inverse regression E(X|y) to a functional data setting. We first estimate the fitted X by regressing X on the vector of length r consisting of centered polynomials of y up to degree r. Then, we compute the functional PFCs (FPFCs). They are obtained by finding a sequence of eigenfunctions of the covariance operator obtained from the standardized fitted X. We add a roughness penalty to the normality constraint imposed on the eigenfunctions to achieve smoothing. The penalty term that controls smoothing is estimated by Generalized Cross Validation. Simulation results highlight the great improvement obtained by FPFCs when it comes to prediction performance compared to FPCs. Furthermore, real data examples show that FPFCs outperform FPCs, especially when data are sparse and longitudinal.

José Luis Torrecilla, Universidad Autónoma de Madrid

Variable selection for classification of functional data

Abstract: The classification of functional data is a significant problem. The variable selection techniques, until now little studied at the functional level, can be effective tools. Their main targets are similar to those of the usual techniques of reducing the dimension (PCA, PLS...) but the variable selection has the advantage of greater interpretability. In this ongoing work we raised several proposals and preliminary results on variable selection in classification taking into account the characteristics of these data, such as redundancy. We explore different ways like using existing algorithms in the multivariate analysis as "minimal redundancy and maximal relevance," mMRM (Ding and Peng 2005), the incorporation of new measures of association for these algorithms and new ways to use these measures based on the study of their maxima.

Javier Gonzalez, University of Groningen

Estimating structured networks using iterative l_1-penalty approaches

Abstract: Many new statistical techniques have been developed in the last years to deal with high dimensional or functional data. The reason is the increasingly improvement of recent measure techniques and the need to analyse the new collected data sets. Some remarkable examples arise in System Biology where the development of new high-throughput tools allows to measure gene expression and protein levels across different time points and experimental conditions.

In the previous context, recent l_1-type regularization methods, such as Graphical lasso or neighbourhood selection models, have been used in Gaussian graphical model selection tasks to recover sparse networks structures from high-dimensional data. However, these approaches do not take into account other inherent properties of real-world networks, i.e. being scale-free and showing modular organization simultaneously.

In this work we propose an alternative iterative l_1-type regularization formulation which is able to reflect in the recovered network the previously mentioned properties. In order to deal with high-dimensional problems in which the data show an inherit functional nature, we use a functional data set-up. The proposed approach can be easy adapted to improve any l_1-type method. In simulation studies we illustrate the performance of our method with respect to some baseline l_1-type approaches. Moreover, we illustrate its behaviour in a real experiment with gene expression time-course data.