MULTIVARIATE ANALYSIS OF COMPLEX DATA WORKSHOP PROGRAM AND ABSTRACTS

May 23rd (Room: R.42.2.113, building R42, Avenue Franklin Roosevelt, 42, 1050, Bruxelles)

- 9:00–9:25: Welcome and coffee
- 9:25-9:30: Short welcome talk
- 9:30–11:00: Session 1 (two speakers)
    1. **Anne Ruiz-Gazen**: "Detecting outliers in compositional data using invariant coordinate selection"
    2. **Valentin Patilea**: "Learning the regularity of curves in functional data analysis and applications"
- 11:00–11:15: Coffee break
- 11:15–12:45: Session 2 (two speakers)
    1. **David Preinerstorfer**: "Consistency of p-norm based tests in high dimensions"
    2. **Richard Samworth**: "Optimal subgroup selection"
- 12:45–13:30: lunch
- 13:30–15:45: Session 3 (three speakers)
    1. **Domenico Marinucci**: "Asymptotics for Spherical Functional Autoregressions"
    2. **Eduardo García-Portugués**: "Tests of uniformity on the hypersphere based on chordal distances"
    3. **Laura Sangalli**: "Spatial and functional data over non-Euclidean domains"
- 15:45–16:00: Coffee break
- 16:00–17:30: Session 4 (two speakers)
    1. **Hongjian Shi**: "On universally consistent and fully distribution-free rank tests of vector independence"
    2. **Gérard Biau**: "A primer on Generative Adversarial Networks"
- 17:30–18:30: Drink in the Atrium

May 24th

- 9:00–9:15: Welcome and coffee
- 9:15–10:45: Session 1 (two speakers)
    1. **Masanobu Taniguchi**: "Joint circular distributions in view of higher order spectra of time series"
    2. **Peter Jupp**: "Ambiguous rotations, inner-product spaces and orientation relationships"
- 10:45–11:00: Coffee break
- 11:00–12:30: Session 2 (two speakers)
    1. **Christophe Ley**: "Multivariate analysis via tools from Stein's Method"
    2. **Rainer von Sachs**: "Statistical inference for intrinsic wavelet estimators of covariance matrices in a log-Euclidean manifold"
- 12:30-12:45: Closing
- 12:45-13:45: Sandwiches

Anne Ruiz-Gazen

Detecting outliers in compositional data using invariant coordinate selection

Invariant Coordinate Selection (ICS) is a multivariate statistical method based on the simultaneous diagonalization of two scatter matrices. A model based approach of ICS, called Invariant Coordinate Analysis, has recently been adapted for compositional data. In a model free context, ICS is also helpful at identifying outliers. We propose to develop a version of ICS for outlier detection in compositional data. This version is first introduced in coordinate space for a specific choice of ilr coordinate system associated to a contrast matrix and follows an existing outlier detection procedure. We then show that the procedure is independent of the choice of contrast matrix and can be defined directly in the simplex. To do so, we first establish some properties of the set of matrices satisfying the zero-sum property and introduce a simplex definition of the Mahalanobis distance and the one-step M-estimators class of scatter matrices. We also define the family of elliptical distributions in the simplex. We then show how to interpret the results directly in the simplex using two artificial datasets and a real dataset of market shares in the automobile industry.

Joint work with Christine Thomas-Agnan (Toulouse School of Economics), Thibault Laurent (Toulouse School of Economics) and Camille Mondon (Ecole Normale Supérieure).

Valentin Patilea

Learning the regularity of curves in functional data analysis and applications

Combining information both within and across trajectories, we propose simple estimators for the local regularity of the trajectories of a stochastic process. Independent trajectories are measured with errors at randomly sampled time points. Non-asymptotic bounds for the concentration of the estimator are derived. Given the estimate of the local regularity, we build a nearly optimal local polynomial smoother from the curves from a new, possibly very large sample of noisy trajectories. We derive non-asymptotic pointwise risk bounds uniformly over the new set of curves. As another application, we build minimax optimal mean and covariance functions estimators. Our estimators perform well in simulations. Real data sets illustrate the effectiveness of the new approaches. The talk is based on joint work with Nicolas Klutchnikoff and Steven Golovkine.

David Preinerstorfer

Consistency of p-norm based tests in high dimensions

Many commonly used test statistics are based on a norm measuring the evidence against the null hypothesis. To understand how the choice of a norm affects power properties of tests in high dimensions, we study the consistency sets of p-norm based tests in the prototypical framework of sequence models with unrestricted parameter spaces, the null hypothesis being that all observations have zero mean. The consistency set of a test is here defined as the set of all arrays of alternatives the test is consistent against as the dimension of the parameter space diverges. We characterize the consistency sets of p-norm based tests. This characterization reveals an unexpected monotonicity result that allows us to construct novel tests that dominate, with respect to their consistency behavior, all p-norm based tests without sacrificing size.

The paper the talk is based on can be downloaded here: https://arxiv.org/abs/2103.11201

Richard Samworth

Optimal subgroup selection

In clinical trials and other applications, we often see regions of the feature space that appear to exhibit interesting behaviour, but it is unclear whether these observed phenomena are reflected at the population level. Focusing on a regression setting, we consider the subgroup selection challenge of identifying a region of the feature space on which the regression function exceeds a pre-determined threshold. We formulate the problem as one of constrained optimisation, where we seek a low-complexity, data-dependent selection set on which, with a guaranteed probability, the regression function is uniformly at least as large as the threshold; subject to this constraint, we would like the region to contain as much mass under the marginal feature distribution as possible. This leads to a natural notion of regret, and our main contribution is to determine the minimax optimal rate for this regret in both the sample size and the Type I error probability. The rate involves a delicate interplay between parameters that control the smoothness of the regression function, as well as exponents that quantify the extent to which the optimal selection set at the population level can be approximated by families of well-behaved subsets. Finally, we expand the scope of our previous results by illustrating how they may be generalised to a treatment and control setting, where interest lies in the heterogeneous treatment effect.

Domenico Marinucci

Asymptotics for Spherical Functional Autoregressions

In this talk, we investigate a class of spherical functional autoregressive processes, and we discuss the estimation of the corresponding autoregressive kernels. In particular, we first establish a consistency result (in sup and mean-square norm), then a quantitative central limit theorem (in Wasserstein distance), and finally a weak convergence result, under more restrictive regularity conditions. We shall also hint at possible extensions, in particular nonparametric testing for stationarity, isotropy and Gaussianity (joint works with Alessia Caponera and Anna Vidotto).

Eduardo Garcia-Portugués

Tests of uniformity on the hypersphere based on chordal distances

We provide a general and tractable family of tests of uniformity on the hypersphere of arbitrary dimension. The family is constructed from powers of the chordal distances between pairs of observations. It connects and extends three particular tests: Rayleigh (1919), Pycke (2007, 2010), and Bakshaev (2010). The asymptotic null distributions of the new tests are obtained and shown to be tractable. Additionally, powers of the tests against generic local alternatives are provided. In particular, explicit powers against Cauchy-like distributions on the hypersphere, that are of independent interest, are derived. Numerical experiments corroborate the obtained theoretical results. Data applications of astronomical and biological nature illustrate the practical use of the tests for assessing uniformity on the two-dimensional sphere.

Laura Sangalli

Spatial and functional data over non-Euclidean domains

Recent years have seen an explosive growth in the recording of increasingly complex and high-dimensional data. Classical statistical methods are often unfit to handle such data, whose analysis calls for the definition of new methods merging ideas and approaches from statistics and applied mathematics. My talk will in particular focus on spatial and functional data defined over non-Euclidean domains, such as linear networks, two-dimensional manifolds and non-convex volumes. I will present an innovative class of methods, based on regularizing terms involving Partial Differential Equations (PDEs), defined over the complex domains being considered. These physics-informed regression methods enable the inclusion of the available problem specific information, suitably encoded in the regularizing PDE. The proposed methods make use of advanced numerical techniques, such as finite element analysis and isogeometric analisys. A challenging application to neuroimaging data will be illustrated.

Hongjian Shi

On universally consistent and fully distribution-free rank tests of vector independence

Rank correlations have found many innovative applications in the last decade. In particular, suitable rank correlations have been used for consistent tests of independence between pairs of random variables. Using ranks is especially appealing for continuous data as tests become distribution-free. However, the traditional concept of ranks relies on ordering data and is, thus, tied to univariate observations. As a result, it has long remained unclear how one may construct distribution-free yet consistent tests of independence between random vectors. This is the problem addressed in this paper, in which we lay out a general framework for designing dependence measures that give tests of multivariate independence that are not only consistent and distribution-free but which we also prove to be statistically efficient. Our framework leverages the recently introduced concept of center-outward ranks and signs, a multivariate generalization of traditional ranks, and adopts a common standard form for dependence measures that encompasses many popular examples. In a unified study, we derive a general asymptotic representation of center-outward rank-based test statistics under independence, extending to the multivariate setting the classical Hájek asymptotic representation results. This representation permits direct calculation of limiting null distributions and facilitates a local power analysis that provides strong support for the center-outward approach by establishing, for the first time, the nontrivial power of center-outward rank-based tests over root-n neighborhoods within the class of quadratic mean differentiable alternatives.

Gérard Biau

A primer on Generative Adversarial Networks

(joint work with B. Cadre, N. Klutchnikoff, A. Stéphanovitch, and U. Tanielian)

The mathematical forces at work behind Generative Adversarial Networks raise challenging theoretical issues. Motivated by the important question of characterizing the geometrical properties of the generated distributions, we provide a thorough analysis of Wasserstein GANs (WGANs) in both the finite sample and asymptotic regimes. We study the specific case where the latent space is univariate and derive results valid regardless of the dimension of the output space. We show in particular that for a fixed sample size, the optimal WGANs are closely linked with connected paths minimizing the sum of the squared Euclidean distances between the sample points. We also highlight the fact that WGANs are able to approach (for the 1-Wasserstein distance) the target distribution as the sample size tends to infinity, at a given convergence rate and provided the family of generative Lipschitz functions grows appropriately. We derive in passing new results on optimal transport theory in the semi-discrete setting.

Masanobu Tanigushi

Joint circular distributions in view of higher order spectra of time series

Circular data analysis is emerging as an important component of statistics. For this half century, various circular distributions have been proposed, e.g., von Mises distribution, wrapped Cauchy distribution, among other things. Also, regarding the joint distribution, Wehrly and Johnson(1980) proposed a bivariate circular distribution which is related to a family of Markov processes on the circle. Because the sample space is on a circle, various new statistical methods have been developed. In this talk we provide a new look at circular distributions in view of spectral distributions of time series because the typical circular distributions correspond to spectral densities of time series models. For example, autoregressive AR(1) spectral density corresponds to wrapped Cauchy distribution, and von Mises distribution corresponds to exponential spectral density (Bloomfield(1973)), etc. Furthermore we introduce a class of joint circular distributions from the higher order spectra of time series, which can describe very general joint circular distributions. Hence we can develop the statistical inference for dependent observations on the circle. We present a family of distributions on the circle derived from the ARMA spectral density. It is seen that the proposed family includes some existing circular families as special cases. For these special cases, the normalizing constant and trigonometric moments are shown to have simple and closed form. We develop the asymptotic optimal inference theory based on the local asymptotic normality (LAN) on the circle. Because the observations are permitted to be dependent, the theory opens a new paradigm in the estimation for joint circular distributions (joint work with Shogo Kato, Institute of Statistical Mathematics,Tokyo, Hiroaki Ogata, Tokyo Metropolitan University and Arthur Pewsey, University of Extremadura, Spain).

Peter Jupp

Ambiguous rotations, inner-product spaces and orientation relationships

The orientation of a physical object in 3-space can be described by a rotation that transforms it into some standard position. For an object (such as a crystal) that is symmetrical, this rotation is known only up to multiplication by an element of the symmetry group, K. Such an ambiguous rotation can be regarded as an element of $SO(3)/K$. A useful tool for handling ambiguous rotations is to embed them into suitable inner-product spaces (a technique used widely in directional statistics). In many crystallographic contexts symmetrical objects occur in pairs and have different symmetry groups. A key concept is that of orientation relationship, a directional form of hidden regression (joint work with Richard Arnold and Helmut Schaeben).

Christophe Ley

Multivariate analysis via tools from Stein's Method

Stein's Method is becoming increasingly popular in statistics and machine learning. In this talk, I will describe how various components from the famous Stein Method, a well-known approach in probability theory for approximation problems, have been recently put to successful use in multivariate analysis.

Rainer von Sachs

Statistical inference for intrinsic wavelet estimators of covariance matrices in a log-Euclidean manifold

In this talk we treat statistical inference for an intrinsic wavelet estimator of curves of symmetric positive definite (SPD) matrices in a log-Euclidean manifold. This estimator preserves positive-definiteness and enjoys permutation-equivariance, which is particularly relevant for covariance matrices. Our second-generation wavelet estimator is based on average-interpolation and allows the same powerful properties, including fast algorithms, known from nonparametric curve estimation with wavelets in standard Euclidean setups.

The core of our work is the proposition of confidence sets for our high-level wavelet estimator in a non-Euclidean geometry. We derive asymptotic normality of this estimator, including explicit expressions of its asymptotic variance. This opens the door for constructing asymptotic confidence regions which we compare with our proposed bootstrap scheme for inference. Detailed numerical simulations confirm the appropriateness of our suggested inference schemes. (Joint work with Johannes Krebs, Eichstaett, and Daniel Rademacher, Heidelberg)