Contents

Abstracts

Interactions of Statistics and Geometry

14–18 February 2022

1 Ming-Yen Cheng

Hong Kong Baptist University, China

Testing specification of distribution in stochastic frontier analysis

Abstract

Stochastic frontier analysis is regularly used in empirical studies to evaluate the productivity and efficiency of companies. A typical stochastic frontier model involves a parametric frontier subject to a composite error term consisting of an inefficiency and a random error. We develop new tests for specification of distribution of the inefficiency. We focus on simultaneous relaxation of two common assumptions: 1) parametric frontier which may lead to false conclusions when misspecified, and 2) homoscedasticity which can be easily violated when working with real data. While these two issues have been extensively studied in prior research exploring the estimation of a stochastic frontier and inefficiencies, they have not been properly addressed in the considered testing problem. We propose novel bootstrap and asymptotic distribution-free tests with neither parametric frontier nor homoscedasticity assumptions, in both cross-sectional and panel settings. Our tests are asymptotically consistent, simple to implement and widely applicable. Their powers against general fixed alternatives tend to one as sample size increases, and they can detect root-n order local alternatives. We demonstrate their efficacies through extensive simulation studies. When applied to a banking panel dataset, our tests provide sound justification for the commonly used exponential specification for banking data. The findings also show that a new parametric frontier model is more plausible than the conventional translog frontier.

2 Ingrid Daubechies

Duke University, USA

Discovering low-dimensional manifolds in high-dimensional data sets

Abstract

Diffusion methods help understand and denoise data sets; when there is additional structure (as is often the case), one can use (and get additional benefit from) a fiber bundle model. This talk reviews diffusion methods to identify low-dimensional manifolds underlying high-dimensional datasets, and illustrates that by pinpointing additional mathematical structure, improved results can be obtained.

Much of the talk draws on a case study from a collaboration with biological morphologists, who compare different phenotypical structures to study relationships of living or extinct animals with their surroundings and each other. This is typically done from carefully defined anatomical correspondence points (landmarks) on e.g. bones; such landmarking draws on highly specialized knowledge. To make possible more extensive use of large (and growing) databases, algorithms are required for automatic morphological correspondence maps, without any preliminary marking of special features or landmarks by the user.

3 Ian Dryden

Florida International University, USA

Non-parametric regression for networks

Abstract

Network data are becoming increasingly available, and so there is a need to develop suitable methodology for statistical analysis. Networks can be represented as graph Laplacian matrices, which are a type of manifold-valued data. Our main objective is to estimate a regression curve from a sample of graph Laplacian matrices conditional on a set of Euclidean covariates, for example in dynamic networks where the covariate is time. We develop an adapted Nadaraya-Watson estimator which has uniform weak consistency for estimation using Euclidean and power Euclidean metrics. We apply the methodology to the Enron email corpus to model smooth trends in monthly

networks and highlight anomalous networks. Another motivating application is given in corpus linguistics, which explores trends in an author's writing style over time based on word co-occurrence networks.

This is joint work with Katie Severn and Simon Preston.

4 Herbert Edelsbrunner

IST Austria, Austria
Distortion, on the average and in expectation

Abstract

We generalize the concept of the Voronoi path of a line to more general shapes and compute the distortion constant, which describes how it changes volume on the average. Although initially asked for a Poisson point process, the distortion is a characteristic property of the space rather than the point process. In other words, the constant ratio of the perimeter of a circle and its pixelation—and the analogous ratios for spheres in three and higher dimensions—hold for all smoothly embedded shapes on average.

This is joint work with Anton Nikitenko.

5 Benjamin Eltzner

Max Planck Institute of Multidisciplinary Sciences, Germany
Testing for uniqueness of estimators

Abstract

Uniqueness of the population descriptor is a standard assumption in asymptotic theory. However, m-estimation problems, in which an estimator is determined by minimizing a cost function, often feature local minima of the sample cost function. These local minima may stem from multiple global minima of the underlying population cost function. We present a hypothesis test to systematically determine for a given sample whether the underlying population cost function may have multiple global minima. The test is widely applicable and we discuss applications to the mean on a non-euclidean data space, nonlinear regression and Gaussian mixture clustering.

6 Aasa Feragen

Technical University of Denmark, Denmark
Predicting graphs

Abstract

Graphs are everywhere! In anatomy and biology, they appear as transportation systems for air, water, nutrients, or signals, and are found both on the large scale of arteries and airways, and on the small scale of neurons in the brain. The structure, geometry and state of the networks affect their function, and therefore also the health of nearby tissue. Conversely, the state of surrounding tissue also affects the networks, making them both first and second order reporters of health, disease and dysfunction. As a consequence, networks are studied extensively in both biology and medicine – and as a proxy for these, in imaging.

In this talk we first discuss a well known space of graphs, where networks are modelled as equivalence classes of adjacency matrices modulo the action of the node permutation group. We derive geometric properties of this space and discuss the implications of those geometric properties for statistics such as dimensionality reduction and graph-valued regression. Next, we discuss the potential for carrying these geometric insights with us into the realm of deep learning on graphs.

7 Sungkyu Jung

Seoul National University, Korea
Clustering on the torus by conformal prediction

Abstract

Motivated by the analysis of torsion (dihedral) angles in the backbone of proteins, we investigate clustering of bivariate angular data on the torus $[-\pi, \pi) \times [-\pi, \pi)$. We show that naive adaptations of clustering methods, designed for vector-valued data, to the torus are not satisfactory and propose a novel clustering approach based on the conformal prediction framework. We construct several prediction sets for toroidal data with guaranteed finitesample validity, based on a kernel density estimate and bivariate von Mises mixture models. From a prediction set built from a Gaussian approximation of the bivariate von Mises mixture, we propose a data-driven choice for the number of clusters and present algorithms for an automated cluster identification and cluster membership assignment. The proposed prediction sets and clustering approaches are applied to the torsion angles extracted from three strains of coronavirus spike glycoproteins (including SARS-CoV-2, contagious in humans). The analysis reveals a potential difference in the clusters of the SARS-CoV-2 torsion angles, compared to the clusters found in torsion angles from two different strains of coronavirus, contagious in animals.

This talk is based on a joint work with B. Kim, K. Park and S. Hong.

8 John Kent

University of Leeds, UK
Statistical methods for semi-concentrated data on manifolds

Abstract

This talk takes a fresh look at the problem of constructing a parametric family of unimodal distributions on a compact manifold. Existing models typically have the flexibility to describe the full range of concentrations (from uniformity at one extreme to a point mass at the other), but are often limited in their ability to describe the full range of eccentricities (from isotropy at one extreme to degeneracy on a submanifold at the other). Distributions which can accommodate the full range of eccentricities are investigated in detail in several settings including the sphere and the torus. Of special interest is the semi-concentrated case, referring to a distribution with low or moderate concentration and high eccentricity, e.g. a unimodal distribution on the sphere concentrated near the equator. Score matching estimation provides an attractive alternative to maximum likelihood estimation which avoids the need to evaluate the normalization constant.

9 Kanti Mardia

University of Leeds and University of Oxford, UK
Statistics of discrete distributions on manifolds: a journey from the Karl Pearson roulette wheel data to some smart health science data

Abstract

Karl Pearson analysed a data from the famous Monte Carlo casino on roulette spins in 1894 which he used as an illustration for his seminal chi-square paper of 1900 – more than one hundred years ago. However, at that time there were no methodologies for analysing directional data so not surprisingly he linearised the problem and constructed a test for unbiasedness of the roulette wheel data on the line. Recently, new discrete circular data are emerging from smart health sciences such as acrophase data. We propose some discrete circular models using marginal and conditionalized approaches to the von Mises distribution and the wrapped Cauchy distribution, and apply these to analyse the data from the two applications areas. These discrete models also provide a benchmark to assess the loss incurred in using any inference from continuous models such as Rayleigh test of uniformity when the underlying circular population is truly discrete. We will also discuss extension of the discrete models to other manifolds such as torus, sphere.

This is a joint work with Dr. Karthik Sriram, Indian Institute of Management Ahmedabad, India.

10 Steve Marron

The University of North Carolina at Chapel Hill, USA
Scaled torus principal component analysis

Abstract

A particularly challenging context for dimensionality reduction is multivariate circular data, i.e., data supported on a torus. Such kind of data appears, e.g., in the analysis of various phenomena in ecology and astronomy, as well as in molecular structures. This paper introduces Scaled Torus Principal Component Analysis (ST-PCA), a novel approach to perform dimensionality reduction with toroidal data. ST-PCA finds a data-driven map from a torus to a sphere of the same dimension and a certain radius. The map is constructed with multidimensional scaling to minimize the discrepancy between pairwise geodesic distances in both spaces. ST-PCA then resorts to principal nested spheres to obtain a nested sequence of subspheres that best fits the data, which can afterwards be inverted back to the torus. Numerical experiments illustrate how ST-PCA can be used to achieve meaningful dimensionality reduction on low-dimensional torii, particularly with the purpose of clusters separation, while two data applications in astronomy (three-dimensional torus) and molecular biology (on a seven-dimensional torus) show that ST-PCA outperforms existing methods for the investigated datasets.

11 Jonathan Mattingly

Duke University, USA

CLTs for empirical measures on stratified spaces

Abstract

I will describe the change in perspective needed to prove that the centered empirical measure converges to a limiting distribution on a riemannian stratified space with curvature bounded from above. This talk will be complementary to the talk by Do Tran, who will describe the structure of the limiting distribution. I will concentrate on understanding how the measure converges.

12 Sayan Mukherjee

Duke University, USA

Modeling shapes and fields: a sheaf theoretic perspective

Abstract

We will consider modeling shapes and fields via topological and lifted-topological transforms. Specifically, we show how the Euler Characteristic Transform and the Lifted Euler Characteristic Transform can be used in practice for statistical analysis of shape and field data. The Lifted Euler Characteristic is an alternative to the. Euler calculus developed by Ghrist and Baryshnikov for real valued functions. We also state a moduli space of shapes for which we can provide a complexity metric for the shapes. We also provide a sheaf theoretic construction of shape space that does not require diffeomorphisms or correspondence. A direct result of this sheaf theoretic construction is that in three dimensions for meshes, 0-dimensional homology is enough to characterize the shape.

13 Hariharan Narayanan

Tata Institute of Fundamental Research, India

Fitting a manifold of large reach to noisy data

Abstract

We give a solution to the following question from manifold learning. Suppose data belonging to a high dimensional Euclidean space is sampled independently, identically at random, from a measure supported on a d dimensional twice differentiable embedded manifold M, and corrupted by Gaussian noise with small standard deviation sigma. How can we produce a manifold M_o whose Hausdorff distance to M is small and whose reach (normal injectivity radius) is not much smaller than the reach of M? We show how to produce a manifold within $O(sigma^2)$ of M in Hausdorf distance, whose reach is smaller than that of M by a factor of no more than $O(d^6)$. This is joint work with Charles Fefferman, Sergei Ivanov and Matti Lassas.

14 Victor M. Panaretos

École polytechnique fédérale de Lausanne (EPFL), Switzerland

The completion of covariance kernels

Abstract

We consider the problem of positive-semidefinite continuation: extending a partially specified covariance kernel from a subdomain of a square domain I x I to a covariance kernel on the entire domain I x I. For a broad class of domains called serrated domains, we will present a complete theory. Namely, we will demonstrate that a canonical completion always exists and can be explicitly constructed. We will characterise all possible completions as suitable perturbations of the canonical completion, and determine necessary and sufficient conditions for a unique completion to exist. We shall interpret the canonical completion via the graphical model structure it induces on the associated Gaussian process. Furthermore, we will show how the estimation of the canonical completion reduces to the solution of a system of linear statistical inverse problems in the space of Hilbert-Schmidt operators, and derive rates of convergence. Time allowing, we will discuss extensions of our theory to more general forms of domains.
(based on joint work with K. Waghmare, EPFL)

15 Wolfgang Polonik

University of California, Davis, USA

Topologically penalized regression on manifolds

Abstract

We study a regression problem on a compact manifold. In order to take advantage of the underlying geometry and topology of the data, we propose to perform the regression task on the basis of eigenfunctions of the Laplace-Beltrami operator of the manifold that are regularized with topological penalties. We will discuss the approach and the penalties, provide some supporting theory and illustrate the performance of the methodology on some data sets. Taken together, our results support the relevance of our approach in the case where the target function is "topologically smooth". This is joint work with O. Hacquard, K. Balasubramanian, G. Blanchard and C. Levrard.

16 Richard Samworth

Cambridge University, UK

Optimal subgroup selection

Abstract

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.

17 Christof Schötz

Universität Heidelberg, Germany

Strong laws of large numbers for Fréchet mean sets

Abstract

A Fréchet mean of a random variable with values in a metric space is an element of the metric space that minimizes the expected squared distance to that random variable. This minimizer may be non-unique. We study strong laws of large numbers for sets of Fréchet means as well as of generalizations where the square is replaced by an arbitrary positive power. We show almost sure convergence of empirical versions of these sets in one-sided Hausdorff distance. The derived results require only minimal assumptions. In particular, only a first moment condition is assumed.

18 Stefan Sommer

University of Copenhagen, Denmark

Stochastic shape analysis and probabilistic geometric statistics

Abstract

Analysis and statistics of shape variation can be formulated in geometric settings with geodesics modelling transitions between shapes. The talk will concern extensions of these smooth geodesic models to account for noise and uncertainty: Stochastic shape processes and stochastic shape matching algorithms. In the stochastic setting, matching algorithms take the form of bridge simulation schemes which also provide approximations of the transition density of the stochastic shape processes. The talk will cover examples of stochastic shape processes and connected bridge simulation algorithms. I will connect these ideas to statistics for data on general manifolds, particularly to the diffusion mean.

19 Do Tran

Georg-August-Universität Göttingen, Germany
CLT of Fréchet mean and geometry

Abstract

We explore relationships between geometry and different forms of CLT, namely *classical*, *smeary* and *sticky*. On Riemannian manifolds, we explain the effect of sectional curvature on behaviors of Fréchet means. In particular, under mild assumptions, Riemannian manifolds which has a section with positive curvature feature smeariness for Fréchet mean. On Riemannian stratified spaces with curvature bounded above, we propose a general form of CLT for probability measures whose support is contained in a ball of radius half the injectivity radius. The general form of the CLT contains information about curvature and singularity of the base space.

This concerns joint works with Stephan Huckemann (University of Goettingen), Ezra Miller (Duke University), and Jonathan Mattingly (Duke University).

20 Katharine Turner

Australian National University, Australia
Computing the extended persistent homology transform of binary images

Abstract

The Persistent Homology Transform, and the Euler Characteristic Transform are topological analogs of the Radon transform that can be used in statsistical shape analysis. In this talk I will consider an interesting variant called the Extended Persistent Homology Transform (XPHT) which replaces the normal persistent homology with extended persistent homology. We are particularly interested in the application of the XPHT to binary images. This paper outlines an algorithm for efficient calculation of the XPHT exploting relationships between the PHT of the boundary curves to the XPHT of the foreground.

21 Andrew Wood

Australian National University, Australia

Score matching for compositional data

Abstract

Compositional data consists of vectors of proportions which are non-negative and sum to 1. Despite involving a familiar space, the simplex, the analysis of compositional data poses a number of challenges and has generated much controversy. A key question is how to accommodate zeros, which often arise in practice. Major limitations of currently available models for compositional data include one or more of the following: insufficient flexibility in terms of distributional shape; difficulty in accommodating zeros in the data in estimation; and lack of computational viability in moderate to high dimensions. A new model for analysing compositional data, the polynomially-tilted pairwise interaction (PPI) model, is discussed. Maximum likelihood estimation is difficult for the PPI model. Instead, we propose novel score matching estimators, which entails extending the score matching approach to Riemannian manifolds with boundary. These new estimators are available in closed form. Theoretical properties of the estimators are discussed and simulation studies shows that these estimators perform well in practice. As our main application we analyse real microbiome count data with fixed totals using a multinomial latent variable model with a PPI model for the latent variable distribution.

This talk is based on the article “Score matching for compositional distributions” by Scealy & Wood (2021, JASA) which has been published online.

22 Ming Yuan

Columbia University, USA

Spectral learning for high dimensional tensors

Abstract

Matrix perturbation bounds developed by Weyl, Davis, Kahan and Wedin and others play a central role in many statistical and machine learning problems. I shall discuss some of the recent progresses in developing similar bounds for higher order tensors. I will highlight the intriguing differences from matrices, and explore their implications in spectral learning problems.