PL01: Ruth Williams

09:30 - 10:30 Thursday, 4th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Howard Bondell

"Stochastic Analysis of Chemical Reaction Networks with Applications to Epigenetic Cell Memory"

DL04: Distinguished Lecture Session

10:50 - 12:30 Thursday, 4th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Hsin-cheng Huang

Yasumasa Matsuda: Fourier analysis of spatio-temporal data

Daisuke Kurisu: Spatially dependent wild bootstrap for high-dimensional spatial data

Yuichi Goto: Integrated copula spectrum with applications to tests for time-reversibility and tail symmetry

DL04.01 Fourier analysis of spatio-temporal data

Your name

Yasumasa Matsuda

Abstract

Fourier analysis has been successfully employed in time series analysis with popular tools of discrete Fourier transform, periodogram, spectral density function that lead to fruitful theoretical and empirical applications in nonparametric or parametric inference. We have interests in extending Fourier analysis of time series to that of spatial/spatio-temporal data, which are principal types of big data collected in the modern age. We define a discrete Fourier transform and periodogram for spatial data and introduce conditions under which they have good asymptotic properties held in time series cases, ie. asymptotic independence. Regarding spatio-temporal data as a time series of functional data, we apply Fourier analysis to identify functional auto-regression in a parametric way, where it should be stressed that they are validated under practical conditions of irregularly spaced samplings, not under those of lattice or continuous observations. Finally, we demonstrate an application to the analysis of relations between human mobility and covid-19 cases collected weekly at 1800 cities in Japan.

DL04.02 Spatially dependent wild bootstrap for high-dimensional spatial data

Your name

Daisuke Kurisu

Abstract

In this paper, we establish a high-dimensional CLT for the sample mean of p-dimensional spatial data observed over irregularly spaced sampling sites in R^d, allowing the dimension p to be much larger than the sample size n. We adopt a stochastic sampling scheme that can generate irregularly spaced sampling sites in a flexible manner and include both pure increasing domain and mixed increasing domain frameworks. To facilitate statistical inference, we develop the spatially dependent wild bootstrap (SDWB) and justify its asymptotic validity in high dimensions by deriving error bounds that hold almost surely conditionally on the stochastic sampling sites. Our dependence conditions on the underlying random field cover a wide class of random fields such as Gaussian random fields and continuous autoregressive moving average random fields. Through numerical simulations and real data analysis, we demonstrate the usefulness of our bootstrap-based inference.

DL04.03 Integrated copula spectrum with applications to tests for time-reversibility and tail symmetry

Your name

Yuichi Goto

Abstract

The spectral density plays a pivotal role in time series analysis. Since the classical spectral density is defined as the Fourier transform of autocovariance functions, it fails to capture the distributional features. To overcome this drawback, we consider the spectral density based on copula and show the weak convergence of integrated copula spectra. This result combined with the subsampling procedure enables us to construct uniform confidence bands, a test for time-reversibility, and a test for tail symmetry. This talk is based on joint work with T. Kley (Georg-August-Univ. Gottingen), R. Van Hecke (Ruhr-Univ. Bochum), S. Volgushev (Univ. of Toronto), H. Dette (Ruhr-Univ. Bochum), and M. Hallin (Univ. libre de Bruxelles).

DL03: Distinguished Lecture Session

10:50 - 12:30 Thursday, 4th January, 2024
EEBrown - Brown Room, Electrical Engineering Building
Chairs Howard Bondell

Jaeyong Lee: Post-processed posteriors for high-dimensional covariances

Kyoungjae Lee: Consistent and scalable Bayesian joint variable and graph selection

Kwangmin Lee: Post-processed posteriors for sparse covariances

DL03.01 Post-processed posteriors for high-dimensional covariances

Your name

Jaeyong Lee

Abstract

We consider Bayesian inference of banded covariance matrices and propose a post-processed posterior. The post-processing of the posterior consists of two steps. In the first step, posterior samples are obtained from the conjugate inverse-Wishart posterior which does not satisfy any structural restrictions. In the second step, the posterior samples are transformed to satisfy the structural restriction through a post-processing function. The conceptually straightforward procedure of the post-processed posterior makes its computation efficient and can render interval estimators of any functional of covariance matrices. We also show that it has nearly optimal minimax rates for banded and bandable covariances among all possible pairs of priors and post-processing functions. The advantages of the post-processed posterior are demonstrated by a simulation study and a real data analysis.

DL03.02 Consistent and scalable Bayesian joint variable and graph selection

Your name

Kyoungjae Lee

Abstract

We consider the joint inference of regression coefficients and the inverse covariance matrix for covariates in high-dimensional probit regression, where the predictors are both relevant to the binary response and functionally related to one another. A hierarchical model with spike and slab priors over regression coefficients and the elements in the inverse covariance matrix is employed to simultaneously perform variable and graph selection. We establish joint selection consistency for both the variable and the underlying graph when the dimension of predictors is allowed to grow much larger than the sample size, which is the first theoretical result in the Bayesian literature. A scalable Gibbs sampler is derived that performs better in high-dimensional simulation studies compared with other state-of-art methods. We illustrate the practical impact and utilities of the proposed method via a functional MRI dataset, where both the regions of interest with altered functional activities and the underlying functional brain network are inferred and integrated together for stratifying disease risk.

DL03.03 Post-processed posteriors for sparse covariances

Your name

Kwangmin Lee

Abstract

We consider Bayesian inference of sparse covariance matrices and propose a post-processed posterior. This method consists of two steps. In the first step, posterior samples are obtained from the conjugate inverse-Wishart posterior without considering the sparse structural assumption. The posterior samples are transformed in the second step to satisfy the sparse structural assumption through a generalized thresholding function. This non-traditional Bayesian procedure is justified by showing that the post-processed posterior attains the optimal minimax rates under the spectral norm loss in high-dimensional settings. We also propose the post-processed posterior for contaminated data and apply it to the estimation of the sparse idiosyncratic covariance of the approximate factor model. The advantages of our method are demonstrated via a simulation study and a real data analysis with S&P 500 data.

IP45: "Recent Advances in Event History Analysis"

10:50 - 12:30 Thursday, 4th January, 2024
B106 - Room 106, Babel Building
Chairs Ting Li
Organiser: (Tony) Jianguo Sun

Ting Li: Conditional Stochastic Interpolation: A New Approach to conditional sampling

Guohao Shen: Nonparametric Estimation of Non-Crossing Quantile Regression Process with Deep ReQU Neural Networks

Kin-Yat Liu: Efficient Estimation for Functional Accelerated Failure Time Model

Xingqiu Zhao: Wasserstein GAN-based Estimation for Conditional Distribution Function with Current Status Data

IP45.01 Conditional Stochastic Interpolation: A New Approach to conditional sampling

Your name

Ting Li

Abstract

We present a novel framework, called the Conditional Stochastic Interpolation, which learns the probability flow equations or stochastic differential equations to transport between two empirically observed distributions. The proposed framework provides a unified solution to conditional generative modeling and domain transfer. The key idea is to learn the velocity function and the conditional score function based on conditional stochastic interpolation, which is then used to construct a Markov process for the purpose of conditional sampling. We derive an upper bound for the excess risk of the estimation of ReLU activated neural networks. To the best of our knowledge, this is the first systematic study on stochastic interpolation with conditions. Numerical studies confirm our theoretical findings. The method is applied to image generating and the aging of human brains, yielding satisfactory performance.

IP45.02 Nonparametric Estimation of Non-Crossing Quantile Regression Processwith Deep ReQU Neural Networks

Your name

Guohao Shen

Abstract

We propose a penalized nonparametric approach to estimating the quantile regression process (QRP) in a nonseparable model using rectifier quadratic unit (ReQU) activated deep neural networks and introduce a novel penalty function to enforce the non-crossing of quantile regression curves. We establish the non-asymptotic excess risk bounds for the estimated QRP and derive the mean integrated squared error for the estimated QRP under mild smoothness and regularity conditions. To establish these non-asymptotic risk and estimation error bounds, we also develop a new error bound for approximating C^s smooth functions with $s >0$ and their derivatives using ReQU activated neural networks. This is a new approximation result for ReQU networks and is of independent interest and may be useful in other problems. Our numerical experiments demonstrate that the proposed method is competitive with or outperforms two existing methods, including methods using reproducing kernels and random forests for nonparametric quantile regression.

IP45.03 Efficient Estimation for Functional Accelerated Failure Time Model

Your name

Liu Kin Yat

Abstract

We propose a functional accelerated failure time model that incorporates both functional and scalar covariates to analyze the time it takes for a particular event to occur. We also establish specific conditions to ensure that the model can be accurately identified. To effectively estimate the model's parameters, we employ a sieve maximum likelihood approach. This approach combines parametric and nonparametric coefficients with an unknown baseline hazard function in the likelihood function. However, this bundling of parameters not only presents significant numerical challenges but also introduces new obstacles in theoretical development. To address these challenges, we develop a comprehensive theoretical framework that allows us to overcome the issues associated with the bundled parameters.

We establish the convergence rate of the proposed estimator, demonstrating that it is consistent as the sample size increases. Additionally, we prove that the finite-dimensional estimator is root-n consistent and is both asymptotically normal and achieves the semiparametric information bound, indicating its desirable statistical properties. To evaluate the effectiveness of our proposed inference procedures, we conduct extensive simulation studies. We also provide an application to illustrate the practical utility of our approach.

IP45.04 Wasserstein GAN-based Estimation for Conditional Distribution Function with Current Status Data

Your name

Xingqiu Zhao

Abstract

Current status data are commonly encountered in modern medicine, econometrics and social science. Its unique characteristics pose significant challenges to the analysis of such data and the existing methods often suffer grave consequences when the underlying model is misspecified. To address these difficulties, we propose a model-free two-stage generative approach for estimating the conditional cumulative distribution function given predictors. We first learn a conditional generator nonparametrically for the joint conditional distribution of observation times and event status, and then construct the nonparametric maximum likelihood estimators of conditional distribution functions based on samples from the conditional generator. Subsequently, we study the convergence properties of the proposed estimator and establish its consistency. Simulation studies under various settings show the superior performance of the deep conditional generative approach over the classical modeling approaches and an application to Parvovirus B19 seroprevalence data yields reasonable predictions.

IP05: "Compositional data analysis: some new fresh approaches which avoid transforming the data"

10:50 - 12:30 Thursday, 4th January, 2024
B305 - Room 305, Babel Building
Chairs Janice Scealy
Organiser: Janice Scealy

Fiona Sammut: Using Generalized Linear Models to Model Compositional Response Data

David Firth: Statistical Analysis of Composition: Principles and Practice

Kassel Hingee: Implementing Score Matching Estimators for Compositional Data and Other Manifold-Valued Data

Janice Scealy: Score matching for microbiome compositional data

IP05.01 Using Generalized Linear Models to Model Compositional Response Data

Your name

Dr Fiona Sammut

Abstract

We develop a multivariate logit model which models the influence of explanatory variables on continuous compositional response variables. This multivariate logit model generalizes an elegant method that was suggested previously by Wedderburn (1974) for the analysis of leaf blotch data in the special case of two categories. In contrast to the logratio modeling approach devised by Aitchison (1982), the multivariate logit model models the expectation of a compositional response variable directly and is also able to handle zeros in the data. The estimation of the parameters in the new model is carried out using the technique of generalized estimating equations (GEE). A working variance-covariance structure which accounts for the specific variability inherent in compositional data, and relevant goodness of fit measures are provided. Properties of the multivariate logit model are also studied empirically using some classic datasets from the literature.

IP05.02 Statistical Analysis of Composition: Principles and Practice

Your name

David Firth

Abstract

The toolbox of compositional data analysis is dominated in applications by log-ratio transformations of multivariate measurements, as advocated in the 1980s by the late John Aitchison based on an axiomatic foundation. In practice the use of log-ratio transformed data often creates difficulties, both interpretational and operational. A particularly well known problem is that zeros in the data make the log-ratio approach inoperable; various fixups have been suggested, using adjusted data, but such adjustments are arbitrary and they inevitably suffer from sensitivity of the logarithm at near-zero values. More modern approaches through statistical models allow much greater flexibility for interpretable analysis across a wide variety of application areas, but typically violate one or more of Aitchison's axioms. In this talk I revisit the Aitchison axioms and assess their implications for, and relevance to, good statistical practice.

IP05.03 Implementing Score Matching Estimators for Compositional Data and Other Manifold-Valued Data

Your name

Kassel Hingee

Abstract

Hyvärinen's score matching estimation is a useful technique for estimating models with intractable normalising constants. However, score matching is sensitive to transformations and the implementation of estimators can involve tedious multivariate differentiation, particularly on bounded manifolds. Quick implementation of estimators for new models and transformations is enabled by automatic differentiation. For compositional data, Janice Scealy, Andrew Wood and John Kent recently found that different transformations lead to different benefits to estimator performance. My score matching package, which uses CppAD automatic differentiation, allows for further exploration of transformations and hybrid estimators. My package can also be used for quick implementation of estimators for models on other manifolds.

IP05.04 Score matching for microbiome compositional data

Your name

Janice Scealy

Abstract

Compositional data are challenging to analyse due to the non-negativity and sum-to-one constraints on the sample space. With real data, it is often the case that many of the compositional components are highly right-skewed, with large numbers of zeros. We propose a new model, the polynomially tilted pairwise interaction (PPI) model, for analysing compositional data. Maximum likelihood estimation is difficult for the PPI model. Instead, we propose two different types of score matching estimators. One is based on extending the score matching approach to Riemannian manifolds with boundary after a square-root transformation. The second approach applies standard score matching estimation after an additive log-ratio transformation. These new estimators are available in closed form and simulation studies show that they perform well in practice. We also define new weighted versions of the estimators and show that they are insensitive to zeros and robust to outliers. An example is given using microbiome data.

IP04: "Statistical Learning"

10:50 - 12:30 Thursday, 4th January, 2024
OA107 - Room 107, Old Arts Building
Chairs Jialiang Li
Organiser: Jialiang Li

Sebastian Kuhnert: Estimating operators of linear and invertible processes in Hilbert spaces, with applications to functional ARMA processes

Jose Angel Sanchez Gomez: Detecting hub variables in large Gaussian graphical models

Mei-Ling Ting Lee: Neural Network for First-hitting-time Based Threshold Regression Model

Wanjie Wang: Network-Adjusted Covariates for Community Detection in Multiscale Networks

---

IP04.01 Estimating operators of linear and invertible processes in Hilbert spaces, with applications to functional ARMA processes

Your name

Sebastian Kühnert

Abstract

Invertible linear processes naturally occur in functional time series analysis. Knowledge of the operators in the linear and inverted representation is of high interest, so consistent operator estimates are of great importance. Explicit asymptotic upper bounds for the full operators in the linear as well as for the finite-dimensional projections and the full operators in the inverted representation of Hilbert space-valued processes have recently been derived. This talk presents the core results of the current article of **Sebastian Kühnert** (UC Davis), Alexander Aue (UC Davis) and Gregory Rice (University of Waterloo). This article derives consistent estimates for the finite-dimensional projections and the full operators under milder conditions, in both the linear and inverted representation of linear, invertible processes in general, separable Hilbert spaces. To be precise, we derive explicit asymptotic upper bounds for the estimation errors under a summability condition, and also exact constants that appear in the consistency results. Moreover, on the basis of these results, we derive novel estimation results for arguably the most important class of functional time series, namely functional MA, AR, and ARMA processes in an arbitrary separable Hilbert space, for the finite-dimensional projections and the full operators, and again with explicit rates and constants.

IP04.02 Detecting hub variables in large Gaussian graphical models

Your name

Jose A. Sanchez Gomez

Abstract

In modern scientific applications, identifying small sets of variables in a dataset with a strong influence over the rest is often vital. For example, when studying the gene-expression levels of cancer patients, estimating the most influential genes can be a first step towards understanding underlying gene dynamics and proposing new treatments. A popular approach for representing variable influence is through a Gaussian graphical model (GGM), where each variable corresponds to a node, and a link between two nodes represents relationships among pairs of variables. In a GGM, influential variables correspond to nodes with a high degree of connectivity, also known as hub variables. In this talk, I share a new method for estimating hub variables in GGMs. To this end, we establish a connection between the presence of hubs in a GGM and the concentration of principal component vectors on the hub variables. We provide probabilistic guarantees of convergence for our method, even in high-dimensional data where the number of variables can be arbitrarily large. I will also discuss an application of this new method to a prostate cancer gene-expression dataset, through which we detect several hub genes with close connections to tumor development.

IP04.03 Neural Network for First-hitting-time Based Threshold Regression Model

Your name

Mei-Ling Ting Lee

Abstract

I will present a neural network expansion (TRNN) of the first hitting time (FHT) based threshold regression (TR) model. The FHT-based TR model conceptualizes a random process for subjects' latent health status. The time-to-event outcome is modeled as the first time the random process hits a threshold. The proposed TRNN can be used in causal survival analysis. The performance of our proposed neural network algorithm is illustrated using real data from a large observational study. TRNN is capable of handling high-dimensional inputs while providing clinically meaningful interpretations.

IP04.04 Network-Adjusted Covariates for Community Detection in Multiscale Networks

Your name

Wanjie Wang

Abstract

Community detection is a crucial task in network analysis that can be significantly improved by incorporating subject-level information, i.e. covariates. Existing methods have shown the effectiveness of using covariates on the low-degree nodes, but rarely discuss the case where communities have significantly different density levels, i.e. multiscale networks.

In this work, we introduce a novel method that addresses this challenge by constructing network-adjusted covariates, which leverage the network connections and covariates with a node-specific weight to each node. This weight can be calculated without tuning parameters.

We present novel theoretical results on the strong consistency of our method under degree-corrected stochastic blockmodels with covariates, even in the presence of mis-specification and multiple sparse communities. Additionally, we establish a general lower bound for the community detection problem when both network and covariates are present, and it shows our method is optimal for connection intensity up to a constant factor.

Our method outperforms existing approaches in simulations and a LastFM app user network. We then compare our method with others on a statistics publication citation network where 30% of nodes are isolated, and our method produces reasonable and balanced

IP40: "Recent Advances in Mathematical Analysis for Data Science"

10:50 - 12:30 Thursday, 4th January, 2024
OA239 - Room 239, Old Arts Building
Chairs Xin Guo
Organiser: Xin Guo

Jun Fan: Generalization analysis of deep ReLU networks for nonlinear functional regression

Yiming Ying: Generalization Analysis for Contrastive Representation Learning

Xin Guo: Capacity dependent analysis for functional online learning algorithms

---

IP40.01 Generalization analysis of deep ReLU networks for nonlinear functional regression

Your name

Jun Fan

Abstract

Neural networks (NNs) are highly versatile tools for approximating continuous functions, but their power goes beyond this. A less known but powerful result is that NNs can also accurately learn nonlinear continuous functionals, making them useful for dynamic system identification, mean field control, and functional data analysis, among other applications. To approximate the functionals defined on Lp spaces, the so-called functional neural networks have been proposed in the literature. However, their theoretical properties are largely unexplored beyond the universality of approximation, and analyses often do not account for the rectified linear unit (ReLU) activation function. In this talk, I will present functional deep ReLU networks and investigate their convergence rates of approximation and generalization errors under different regularity conditions.

IP40.02 **Generalization Analysis for Contrastive Representation Learning**

Your name

Yiming Ying

Abstract

The performance of machine learning (ML) models often depends on the representation of data, which motivates a resurgence of contrastive representation learning (CRL) to learn a representation function. Recently, CRL has shown remarkable empirical performance and it can even surpass the performance of supervised learning models in various domains such as computer vision and natural language processing. In this talk, I present our recent progress in establishing the learning theory foundation for CRL. In particular, we address the following two theoretical questions: 1) how would the generalization behavior of downstream ML models benefit from the representation function built from positive and negative pairs? 2) Especially, how would the number of negative examples affect its learning performance? Specifically, we can show that generalization bounds for contrastive learning do not depend on the number k of negative examples, up to logarithmic terms. Our analysis uses structural results on empirical covering numbers and Rademacher complexities to exploit the Lipschitz continuity of loss functions. For self-bounding Lipschitz loss functions, we further improve our results by developing optimistic bounds which imply fast rates in a low noise condition. We apply our results to learning with both linear representation and nonlinear representation by deep neural networks, for both of which we derive explicit Rademacher complexity bounds.

IP40.03 **Capacity dependent analysis for functional online learning algorithms**

Your name

Xin Guo

Abstract

This research provides convergence analysis of online stochastic gradient descent algorithms for functional linear models. Adopting the characterizations of the slope function regularity, the kernel space capacity, and the capacity of the sampling process covariance operator, significant improvement in the convergence rates is achieved. Both prediction problems and estimation problems are studied, where we show that capacity assumption can alleviate the saturation of the convergence rate as the regularity of the target function increases.

We show that with a properly selected kernel, capacity assumptions can fully compensate for the regularity assumptions for prediction problems (but not for estimation problems). This demonstrates the significant difference between the prediction problems and the estimation problems in functional data analysis.

IP54: "Analysis of data evolving with time"

10:50 - 12:30 Thursday, 4th January, 2024
EEBrown - Brown Room, Electrical Engineering Building
Chairs Aurore Delaigle
Organiser: Aurore Delaigle

Martin Hazelton: Shrinkage estimators of the spatial relative risk function

Degui Li: Estimating Covariance Functions for High-Dimensional Functional Time Series with Dual Factor Structures

Rob Hyndman: Forecast reconciliation: a brief overview

Jeroen Rombouts: Monitoring Machine Learning Forecasts for Platform Data Streams

IP54.01 Shrinkage estimators of the spatial relative risk function

Your name

Martin Hazelton

Abstract

The spatial relative risk function describes differences in the geographical distribution of two types of points, such as locations of cases and controls in an epidemiological study. It is defined as the ratio of the two underlying densities. Estimation of the spatial relative risk function is usually done using kernel estimates of these densities. However, this procedure is often challenging in practice because of the high degree of spatial inhomogeneity in the distributions. This makes it difficult to obtain estimates of the relative risk that are stable in areas of sparse data while retaining necessary detail elsewhere, and consequently difficult to distinguish true risk hotspots from stochastic bumps in the risk function. We study shrinkage estimators of the spatial relative risk function to address these problems. In particular, we propose a new lasso-type estimator that shrinks a standard kernel estimator of the log-relative risk function towards zero, eliminating stochastic bumps.

IP54.02 Estimating Covariance Functions for High-DimensionalFunctional Time Series with Dual Factor Structures

Your name

Degui Li

Abstract

We consider high-dimensional stationary functional time series with dual functional factor model structures. A high-dimensional fully functional factor model is imposed on the observed functional processes whereas a low-dimensional one is assumed for the latent functional factors via the series approximation. Our primary interest is to estimate the large matrix of covariance functions which satisfies the so-called functional “low-rank plus sparse” structure. We extend the classic principal component analysis technique to functional time series and estimate the functional factor loadings, common factors and covariance of functional common components. A functional generalised shrinkage is subsequently applied to the estimated idiosyncratic covariance functions. Under some regularity conditions, we derive the large sample theory of the developed estimates, including the consistency of the estimated factors and functional factor loadings and the uniform convergence rates of the estimated covariance functions. Both the simulation and empirical studies are provided to demonstrate reliable finite-sample performance of the developed model and estimation methodology.

IP54.03 Forecast reconciliation: a brief overview

Your name

Rob J Hyndman

Abstract

Collections of time series that are formed via aggregation are prevalent in many fields. These are commonly referred to as hierarchical time series and may be constructed cross-sectionally across different variables, temporally by aggregating a single series at different frequencies, or may even be generalised beyond aggregation as time series that respect linear constraints. When forecasting such time series, a desirable condition is for forecasts to be coherent, that is to respect the constraints. The past decades have seen substantial growth in this field with the development of reconciliation methods that not only ensure coherent forecasts but can also improve forecast accuracy. This talk provides an overview of recent work on forecast reconciliation.

IP54.04 Monitoring Machine Learning Forecasts for Platform Data Streams

Your name

Jeroen Rombouts

Abstract

Data stream forecasts are essential inputs for decision making at digital platforms. Machine learning (ML) algorithms are appealing candidates to produce such forecasts. Yet, digital platforms require a large-scale forecast framework that can flexibly respond to sudden performance drops. Re-training ML algorithms at the same speed as new data batches enter is usually computationally too costly. On the other hand, infrequent re-training requires specifying the re-training frequency and typically comes with a severe cost of forecast deterioration. To ensure accurate and stable forecasts, we propose a simple data-driven monitoring procedure to answer the question when the ML algorithm should be re-trained. Instead of investigating instability of the data streams, we test if the incoming streaming forecast loss batch differs from a well-defined reference batch. Using a novel dataset constituting 15-min frequency data streams from an on-demand logistics platform operating in London, we apply the monitoring procedure to popular ML algorithms including random forest, XGBoost and lasso. We show that monitor-based re-training produces accurate forecasts compared to viable benchmarks while preserving computational feasibility. Moreover, the choice of monitoring procedure is more important than the choice of ML algorithm, thereby permitting practitioners to combine the proposed monitoring procedure with one's favorite forecasting algorithm.

CP07: Contributed Paper Session

10:50 - 12:30 Thursday, 4th January, 2024
BG03 - Room G03, Babel Building
Chairs Subhrajyoty Roy

Subhrajyoty Roy: rSVDdpd: A Robust Scalable Video Surveillance Background Modelling Algorithm

Takuya Koriyama: Correcting generalized cross-validation for arbitrary ensembles of penalized estimators

Satish Kumar: A novel characterization of structures in smooth regression curves: from a viewpoint of persistent homology

CP07.02 rSVDdpd: A Robust Scalable Video Surveillance Background Modelling Algorithm

Your name

Subhrajyoty Roy

Abstract

In automated video surveillance, a basic algorithmic task involves detecting background and foreground contents, with numerous applications in defence, security, research and monitoring, etc. Camera tampering, noisy videos, low frame rates, etc. pose challenges in solving the problem. The traditional method of classifying the tampered frames, and conducting subsequent analysis on non-tampered frames, often leads to information loss. Although existing methods based on robust principal component analysis (RPCA) address this issue, they are computationally expensive. On the other hand, emergent deep-learning methods are faster, but they demand large amounts of training data and sophisticated hardware to perform this task.

In this paper, we introduce a novel robust singular value decomposition technique (rSVDdpd) based on the popular minimum density power divergence estimator, to solve the video surveillance background modelling problem robustly. Notably, the proposed approach is unsupervised, requires no training data, and can be implemented with little computational cost by using a simple iterative technique based on alternating weighted regression. We also established theoretical guarantees like equivariance, convergence, consistency, etc. of the rSVDdpd estimator. To validate its efficacy, we compare its performance against several existing RPCA algorithms on a benchmark dataset and a new “University of Houston Camera Tampering Dataset”.

This work is an application-driven one that focuses on a particular problem of practical interest, that of video surveillance background modelling. The main theoretical component of the proposed method, robust singular value decomposition based on density power divergence, has many other potential applications.

CP07.03 Correcting generalized cross-validation for arbitrary ensembles of penalized estimators

Your name

Takuya Koriyama

Abstract

Generalized cross-validation (GCV) employs a scalar data-dependent degrees of freedom adjustment to the squared training error and serves as an extremely fast method for estimating the squared out-of-sample prediction risk without needing any sample splitting or any model refitting. In this paper, we examine consistency of GCV for estimating the prediction risk of arbitrary ensembles of penalized estimators. We show that for any finite ensemble, GCV is inconsistent. Towards repairing this shortcoming, we identify a correction to GCV (termed CGCV), which involves an additional scalar data-dependent correction based on degrees of freedom adjusted training errors of the individual component predictors. CGCV maintains all the computational advantages of GCV, and in particular, requires neither any sample splitting nor any model refitting nor any out-of-bag risk estimation. The structure CGCV stems from a finer inspection of ensemble risk decomposition, and two intermediate risk estimators for the components in this decomposition. We provide a non-asymptotic analysis of the CGCV and the two intermediate risk estimators for ensembles of convex penalized estimators under Gaussian features and a linear response model. Furthermore, in the special case of ridge regression, we extend the analysis for general feature and response distributions that assume only mild moment bounds.

CP07.04 A novel characterization of structures in smooth regression curves: from a viewpoint of persistent homology.

Your name

Satish Kumar

Abstract

In this talk, I shall talk about a novel characterization of structures such as monotonicity, convexity, and modality in smooth regression curves using persistent homology. Persistent homology is a key tool in topological data analysis that detects higher dimensional topological features such as connected components and holes (cycles or loops) in the data. Persistent homology is a multiscale version of homology that characterizes sets based on the connected components and holes. We use super-level sets of functions to extract geometric features via persistent homology. In particular, we explore structures in regression curves via the persistent homology of super-level sets of a function, where the function of interest is - the first derivative of the regression function. In the course of this study, we extend an existing procedure of estimating the persistent homology for the first derivative of a regression function and establish its consistency. Moreover, as an application of the proposed methodology, we demonstrate that the persistent homology of the derivative of a function can reveal hidden structures in the function that are not visible from the persistent homology of the function itself. In addition, we also illustrate that the proposed procedure can be used to compare the shapes of two or more regression curves which is not possible merely from the persistent homology of the function itself.

CP07.05 Bayesian sample size determination for network structure learning

Your name

Guido Consonni

Abstract

Graphical models based on Directed Acyclic Graphs (DAGs) are widely used to model dependence relations among variables. Observational data cannot distinguish in general between DAGs representing the same conditional independence assertions (Markov equivalent DAGs), which implies a limited ability to learn causal relationships among variables. Interventional data produced after exogenous manipulations of the variables can greatly improve the structure learning process, because they can distinguish among equivalent DAGs.

Since interventions are typically expensive, an optimal design aiming at minimizing the number of variables to be manipulated is crucial. Most importantly however, the accuracy behind the DAG learning process depends on the number of collected interventional data, whose sample size should be determined before the intervention experiment is performed.

We tackle this problem from a Bayesian experimental design perspective by adopting the Bayes Factor as a measure of evidence between competing causal structures. For any candidate sequence of manipulated variables, our method determines the corresponding optimal sample size which guarantees a fixed degree of assurance that the intervention experiment will produce compelling evidence in favor of the true causal-model hypothesis.

DL09: Distinguished Lecture Session

13:30 - 15:10 Thursday, 4th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Fang Yao

Jane-Ling Wang: The Many faces of Functional Data: The Trouble with Sparse Functional Data

Hans-Georg Mueller: Metric Statistics: Exploration of Random Objects With Distance Profiles

Kuang-Yao Lee: Nonparametric functional graphical models

DL09.01 The Many faces of Functional Data: The Trouble with Sparse Functional Data

Your name

Jane-Ling Wang

Abstract

Functional data are random functions on an interval, e.g. [0, 1]. They have become increasingly common due to advances in modern technology to collect and store such data. In reality these random functions can only be measured at discrete time grids and the measurement schedule may vary among subjects. Depending on the sampling frequency, functional data are collected either intensively or sparsely, which affects both methodology and theory. Furthermore, the data may contain noise, a.k.a. measurement errors. Among the various sampling plans, sparsely observed functional data that feature only a few measurements per subject are the most challenging type to deal with, both in methodology and theory. Although the challenges for mean and covariance estimation and many regression settings have been addressed for such sparse functional data, many other tasks are infeasible with sparsely observed functional data. In this talk, we discuss several examples, where consistent inference is infeasible for sparsely observed functional data and show what can be accomplished in each case. For instance, we show that testing the equality of distributions, a.k.a. homogeneity test, between two independent samples of functional data is infeasible for sparse functional data but a test of marginal homogeneity, i.e., the marginal distributions are the same among the independent samples for all time points t, is attainable. The talk will end with a brief discussion of a current direction that extends approaches for functional data to data in a metric space, which may not be a linear space.

DL09.02 Metric Statistics: Exploration of Random Objects With Distance Profiles

Your name

Hans-Georg Müller

Abstract

For the statistical modeling of complex data, which are increasingly encountered in modern data analysis, a general approach is to view such data as elements of a metric space that satisfies certain structural conditions and features a probability measure. We refer to the random elements of such spaces as random objects and to the emerging field that deals with their statistical analysis as metric statistics. A focus of this talk are distance profiles as a major tool for the analysis of samples of random objects in conjunction with the pairwise Wasserstein transports of distance profiles which are one-dimensional distance distributions. These pairwise transports give rise to an alternative transport metric that complements the original metric of the object space and is shown to reveal important features of the object data, which include intuitive and interpretable notions of transport ranks and transport quantiles.The convergence of sample based estimates to their population targets provides theoretical justifications. Specific examples for the utility and visualization of distance profiles and transport ranks include distributional data, compositional data and network data. This talk is based on joint work with Yaqing Chen (Rutgers) and Paromita Dubey (USC).

DL09.03 Nonparametric functional graphical models

Your name

Kuang-Yao Lee

Abstract

In this work we develop a nonparametric graphical model for multivariate random functions. Most existing graphical models are restricted by the assumptions of multivariate Gaussian or copula Gaussian distributions, which also imply linear relations among the random variables or functions on different nodes. We relax those assumptions by building our graphical model based on a new statistical object—the functional additive regression operator. By carrying out regression and neighborhood selection at the operator level, our method can capture nonlinear relations without requiring any distributional assumptions. Moreover, the method is built up using only one-dimensional kernel, thus, avoids the curse of dimensionality from which a fully nonparametric approach often suffers, and enables us to work with large-scale networks.

We derive error bounds for the estimated regression operator and establish graph estimation consistency, while allowing the number of functions to diverge at the exponential rate of the sample size. We demonstrate the efficacy of our method by both simulations and analysis of an electroencephalography dataset. (This is joint work with Lexin Li (UC Berkeley), Bing Li (PSU), and Hongyu Zhao (Yale))

DL01: Distinguished Lecture Session

13:30 - 15:10 Thursday, 4th January, 2024
B106 - Room 106, Babel Building
Chairs Minge Xie

Siva Athreya: Path convergence of Markov chains on large graphs

Qunqiang Feng: Average Jaccard index of random graphs

Qingwei Liu: Normal approximation of subgraph counts in the random-connection model

---

DL01.01 Path convergence of Markov chains on large graphs

Your name

Siva Athreya

Abstract

We consider two classes of natural stochastic processes on finite unlabeled graphs: Euclidean stochastic optimization algorithms on the adjacency matrix of weighted graphs and a modified version of the Metropolis MCMC algorithm on stochastic block models over unweighted graphs. In both cases we show that, as the size of the graph goes to infinity, the random trajectories of the stochastic processes converge to deterministic curves on the space of measure-valued graphons. Measure-valued graphons, introduced by Lov\'{a}sz and Szegedy in \cite{lovasz2010decorated}, are a refinement of the concept of graphons that can distinguish between two infinite exchangeable arrays that give rise to the same graphon limit. We introduce new metrics on this space which provide us with a natural notion of convergence for our limit theorems. This notion is equivalent to the convergence of infinite-exchangeable arrays. Under suitable assumptions and a specified time-scaling, the Metropolis chain admits a diffusion limit as the number of vertices go to infinity. We then demonstrate that, in an appropriately formulated zero-noise limit, the stochastic process of adjacency matrices of this diffusion converges to a deterministic gradient flow curve on the space of graphons introduced in~\cite{Oh2023}. A novel feature of this approach is that it provides a precise exponential convergence rate for the Metropolis chain in a certain limiting regime. The connection between a natural Metropolis chain commonly used in exponential random graph models and gradient flows on graphons, to the best of our knowledge, is also new in the literature.

DL01.02 Average Jaccard index of random graphs

Your name

Qunqiang Feng

Abstract

The asymptotic behavior of the Jaccard index in G(n, p), the classical Erdos-Renyi random graphs model, is studied in this paper, as n goes to infinity. We first derive the asymptotic distribution of the Jaccard index of any pair of distinct vertices, as well as the first two moments of this index. Then the average of Jaccard indices over all vertex pairs in G(n, p) is shown to be asymptotically normal under an additional mild condition.

DL01.03 Normal approximation of subgraph counts in the random-connection model

Your name

Qingwei Liu

Abstract

We derive normal approximation results for subgraph counts written as multiparameter stochastic integrals in a random-connection model based on a Poisson point process. By combinatorial arguments we express the cumulants of general subgraph counts using sums over connected partition diagrams, after cancellation of terms obtained by M˙obius inversion. Using the Statuleviˇcius condition, we deduce convergence rates in the Kolmogorov distance by studying the growth of subgraph count cumulants as the intensity of the underlying Poisson point process tends to infinity. Our analysis covers general subgraphs in the dilute and full random graph regimes, and tree-like subgraphs in the sparse random graph regime.

IP01: "Non-normal approximations and their applications"

13:30 - 15:10 Thursday, 4th January, 2024
B305 - Room 305, Babel Building
Chairs Nathan Ross
Organiser: Lihu Xu

Xiao Fang: High-dimensional Central Limit Theorems by Stein's Method in the Degenerate Case

Adrian Rollin: Statistical Applications of Centred Subgraph Counts in Statistical Network Analysis

Nathan Ross: Gaussian random field approximation for wide neural networks

Songhao Liu: General Non-normal Approximation With Unbounded Exchangeable Pairs

IP01.01 High-dimensional Central Limit Theorems by Stein's Method in the Degenerate Case

Your name

Xiao Fang

Abstract

In the literature of high-dimensional central limit theorems, there is a gap between results for general limiting correlation matrix Σ and the strongly non-degenerate case. For the general case where Σ may be degenerate, under certain light-tail conditions, when approximating a normalized sum of n independent random vectors by the Gaussian distribution $N(0,\Sigma)$ in multivariate Kolmogorov distance, the best-known error rate has been $O(n^{-1/4})$, subject to logarithmic factors of the dimension. For the strongly non-degenerate case, that is, when the minimum eigenvalue of Σ is bounded away from 0, the error rate can be improved to $O(n^{-1/2})$ up to a $\log n$ factor. In this paper, we show that the $O(n^{-1/2})$ rate up to a $\log n$ factor can still be achieved in the degenerate case, provided that the minimum eigenvalue of the limiting correlation matrix of any three components is bounded away from 0. We prove our main results using Stein's method in conjunction with previously unexplored inequalities for the integral of the first three derivatives of the standard Gaussian density over convex polytopes. These inequalities were previously known only for hyperrectangles. Our proof demonstrates the connection between the three-components condition and the third moment Berry--Esseen bound.

IP01.02 Statistical Applications of Centred Subgraph Counts in Statistical Network Analysis

Your name

Adrian Röllin

Abstract

We discuss the theory and application of so-called centred subgraph counts to the statistical analysis of complex networks, in particular to goodness-of-fit analysis. The theory is based on higher-order fluctuation theory developed in the 90s by Janson and Nowicki to understand the distributional behaviour of generalised U-statistics.

IP01.03 Gaussian random field approximation for wide neural networks

Your name

Nathan Ross

Abstract

It has been observed that wide neural networks (NNs) with randomly initialized weights may be well-approximated by Gaussian fields indexed by the input space of the NN, and taking values in the output space. There has been a flurry of recent work making this observation precise, since it sheds light on regimes where neural networks can perform effectively. In this talk, I will discuss recent work where we derive bounds on Gaussian random field approximation of wide random neural networks of any depth. The bounds are on a Wasserstein transport distance in function space equipped with a strong (supremum) metric, and are explicit in the widths of the layers and natural parameters such as moments of the weights. The result follows from a general approximation result using Stein's method, combined with a novel Gaussian smoothing technique for random fields.

The talk covers joint works with Krishnakumar Balasubramanian, Larry Goldstein, and Adil Salim; and A.D. Barbour and Guangqu Zheng.

IP01.04 General Non-normal Approximation With Unbounded Exchangeable Pairs

Your name

Liu, Song-Hao

Abstract

Using exchangeable pairs approach of Stein's method, we develop a general non-normal Berry-Esseen bound for general unbounded exchangeable pairs. The result extends results of Shao and Zhang (2019). Applications to Pearson's statistic, Polya urn model and isotropic mean-field Heisenberg model are also discussed.

IP56: "Recent developments in Survival Analysis and Statistical Machine Learning"

13:30 - 15:10 Thursday, 4th January, 2024
AW153 - Room 153, Arts West North Building
Chairs Xingqiu Zhao
Organiser: Xingqiu Zhao

Qiuzhuang Sun: Optimal stopping with partially observable information

Kin-Yau Wong: Improving estimation efficiency for additive hazard models with varying coefficients

Yuchen Wu: Posterior Sampling from the Spiked Models via Diffusion Processes

Wen Su: Deep Nonparametric Inference for Conditional Hazard Function

IP56.01 Optimal stopping with partially observable information

Your name

Qiuzhuang Sun

Abstract

We study an optimal stopping problem with partial information and apply the method to mission-critical systems, which are required to operate continuously to complete a mission. When there is a sign of imminent failures, the mission can be aborted to increase the system survival probability and minimize damage to the system. We consider the abort (optimal stopping) decision-making problem for systems with three states: healthy, defective, and failure, where healthy and defective states are unobservable. Condition-monitoring sensors are installed and periodically generate signals indicating the system health state. However, possible measurement errors result in imperfect sensitivity and specificity of the sensors, significantly complicating the decision-making. Furthermore, the random time from system defect to failure typically has an increasing hazard rate, leading to a non-Markovian system deterioration process. This study meets these challenges by adopting the Erlang mixture distributions to approximate the non-Markovian failure process as a continuous-time Markov chain in a new state space. A partially observable Markov decision process (POMDP) is then formulated for decision-making. We show that the optimal policy of the POMDP follows a control-limit structure in a spherical coordinate system, and the sequence of optimal policies converges to the true optimal policy for the original problem when the phase of the Erlang mixture goes to infinity. A modified point-based value iteration algorithm is developed to deal with the curse of dimensionality. We further investigate two special cases of our model that can be exactly solved after discretizing the state space.

IP56.02 Improving estimation efficiency for additive hazard models with varying coefficients

Your name

Kin Yau Wong

Abstract

In contemporary omics studies of chronic diseases, investigators are often interested in the association between multiple types of omics/clinical features and outcomes of interest, such as the time to disease progression or death. In most existing models, effects of individual features are assumed to be independent of other omics features and clinical variables, that is, interaction effects are not considered. In this project, we consider varying-coefficient additive hazards models that allow for effects of a type of features to be modified by other omics or clinical features. We will develop a novel estimation approach based on kernel smoothing, which is more efficient than existing methods. We establish the asymptotic properties of the proposed estimators and demonstrate the feasibility and advantages of the proposed methods through extensive simulation studies and real data analyses.

IP56.03 Posterior Sampling from the Spiked Models via Diffusion Processes

Your name

Yuchen Wu

Abstract

Sampling from the posterior is a key technical problem in Bayesian statistics. Rigorous guarantees are difficult to obtain for Markov Chain Monte Carlo algorithms of common use. In this paper, we study an alternative class of algorithms based on diffusion processes. The diffusion is constructed in such a way that, at its final time, it approximates the target posterior distribution. The stochastic differential equation that defines this process is discretized (using a Euler scheme) to provide an efficient sampling algorithm. Our construction of the diffusion is based on the notion of observation process and the related idea of stochastic localization. Namely, the diffusion process describes a sample that is conditioned on increasing information. An overlapping family of processes was derived in the machine learning literature via time-reversal.

We apply this method to posterior sampling in the high-dimensional symmetric spiked model. We observe a rank-one matrix θθT corrupted by Gaussian noise, and want to sample θ from the posterior. Our sampling algorithm makes use of an oracle that computes the posterior expectation of θ given the data and the additional observation process. We provide an efficient implementation of this oracle using approximate message passing. We thus develop the first sampling algorithm for this problem with approximation guarantees.

IP56.04 Deep Nonparametric Inference for Conditional Hazard Function

Your name

Wen Su

Abstract

We propose a novel deep learning approach to nonparametric statistical inference for the conditional hazard function of survival time with right-censored data. We use a deep neural network (DNN) to approximate the logarithm of a conditional hazard function given covariates and obtain a DNN likelihood-based estimator of the conditional hazard function. Such an estimation approach grants model flexibility and hence relaxes structural and functional assumptions on conditional hazard or survival functions. We establish the consistency, convergence rate, and functional asymptotic normality of the proposed estimator. Subsequently, we develop new one-sample tests for goodness-of-fit evaluation and two-sample tests for treatment comparison. Both simulation studies and real application analysis show superior performances of the proposed estimators and tests in comparison with existing methods.

IP08: "Information Geometry and its Related Fields"

13:30 - 15:10 Thursday, 4th January, 2024
OA239 - Room 239, Old Arts Building
Chairs Tomonari Sei
Organiser: Tomonari Sei

Hiroshi Matsuzoe: Invariant dually flat structures on deformed exponential families

Tomonari Sei: Minimum information dependence model and its conditional inference

Yoshihiro Hirose: LARS Algorithm in Tangent Space of Generalized Linear Models

Takeru Matsuda: Wasserstein--Cramer--Rao inequality and robustness

IP08 Invariant dually flat structures on deformed exponential families

Your name

Hiroshi Matsuzoe

Abstract

In information geometry, there are two important concepts for understanding the statistical manifold of probability density functions: invariance and dual flatness. Invariance is the property that the geometric structure of the statistical manifold formed by the family of probability density functions does not change under the transformation of random variables. Dual flatness is the property that the Riemannian metric of a statistical manifold is not Euclidean, but the dual affine connections are flat. It is known that an exponential family is a dually flat space. The Fisher metric is a Hessian metric and the dual coordinate system with respect to a natural coordinate system can be obtained from the Legendre transformation.

Information geometry of deformed exponential families has focused mainly on dual flatness. Different from the ordinary exponential family, there are several natural dually flat structures in deformed exponential families. In this talk, we consider invariant and dually flat geometric structures in deformed exponential families. Since this structure cannot be calculated directly from the non-additive entropy, a new construction method for dually flat space will be given.

IP08.02 Minimum information dependence model and its conditional inference

Your name

Tomonari Sei

Abstract

We propose a multivariate model called the minimum information dependence model. The model is characterized by two orthogonal sets of parameters: the dependence parameter and the marginal parameter. The parameterization is a mixed coordinate system in the framework of information geometry. The multivariate Gaussian model for quantitative data and log-linear models for categorical data are particular cases of our model. A conditional inference method given the marginal empirical distribution is shown to be useful. Some illustrative examples are provided.

This is a joint work with Keisuke Yano (The Institute of Statistical Mathematics).

IP08.03 LARS Algorithm in Tangent Space of Generalized Linear Models

Your name

Yoshihiro Hirose

Abstract

Least angle regression (LARS) is one of the famous sparse estimation algorithms for linear models. In this talk, the LARS algorithm is applied to the tangent space of generalized linear models. The space of the normal linear model is a Euclidean space while that of a generalized linear model is a manifold, not a Euclidean space. One way to apply the LARS algorithm to the generalized linear models is to extend the algorithm such that it works in a manifold. However, in this talk, we adopt another approach. That is, we approximate the manifold of the generalized linear model with its tangent space, and the LARS algorithm is applied to that tangent space. This approach leads to efficient computation of the method because the LARS algorithm is easy to compute. In this talk, the estimation method is explained, and numerical evaluation is also presented.

IP08.04 Wasserstein--Cramer--Rao inequality and robustness

Your name

Takeru Matsuda

Abstract

Wasserstein distance is the optimal transportation cost between probability distributions. It induces its own geometric structure on the set of probability distributions, which is different from the information geometric structure induced by the Kullback-Leibler divergence. Recently, Li and Zhao (2023) developed Wasserstein counterparts of information geometric concepts such as Wasserstein information matrix, Wasserstein score and Wasserstein--Cramer--Rao inequality. In this study, we consider the Wasserstein geometry of elliptically contoured families and show that the Wasserstein covariance quantifies the robustness against additive noise of quadratic estimators and the Wasserstein--Cramer--Rao inequality gives its lower bound.

IP15: "Inference for partially observed structured dynamic systems"

13:30 - 15:10 Thursday, 4th January, 2024
OA107 - Room 107, Old Arts Building
Chairs Edward Ionides
Organiser: Edward Ionides

Patricia (Ning) Ning: Variable Target Scalable Particle Filter

Kevin Tan: Automatic Differentiation to Accelerate Likelihood-Based Inference on Partially Observed Stochastic Processes

Edward Ionides: An iterated block particle filter for inference on coupled dynamic systems

Carles Breto: Panel data analysis via mechanistic models

IP15.01 Variable Target Scalable Particle Filter

Your name

Ning Ning

Abstract

We address the challenge of tracking a variable number of interacting targets. Our primary goal is to accurately detect targets entering and leaving the scene while maintaining a precise trajectory record for each target throughout their presence. Tracking multiple indistinguishable targets that interact with one another presents a formidable challenge. Therefore, developing methods that effectively handle this complexity is vital for scenarios involving the continuous tracking of numerous interacting targets. To tackle this problem, we introduce the Variable Target Scalable Particle Filter (VTSPF) within an online learning framework. VTSPF efficiently tracks multiple moving targets exhibiting complex interactions. Importantly, it demonstrates scalability in both spatial and temporal dimensions.

IP15.02 Accelerated Inference for Partially Observed Stochastic Processes with Automatic Differentiation

Your name

Kevin Tan

Abstract

Automatic differentiation (AD) has driven recent advances in machine learning, including deep neural networks and Hamiltonian Markov Chain Monte Carlo methods. Partially observed nonlinear stochastic dynamical systems have proved resistant to AD techniques due to (1) the requirement for simulation-based inference that does not require access to the system's transition probabilities and (2) the issue that widely used particle filter algorithms yield a likelihood function that is discontinuous as a function of the model parameters. We propose a new representation that embeds existing methods in a theoretical framework that readily permits extension to a new class of algorithms, providing opportunities for optimizing a bias/variance tradeoff. Further, we develop optimization algorithms suited to the Monte Carlo properties of the derivative estimate, including a hybrid algorithm that requires only a differentiable simulator for maximum likelihood estimation. Promising numerical results indicate that a hybrid algorithm that uses AD to refine a coarse solution from iterated filtering can beat current state-of-the-art methods on a challenging scientific benchmark problem.

IP15.03 An iterated block particle filter for inference on coupled dynamic systems

Your name

Edward Ionides

Abstract

We consider inference for a collection of partially observed, stochastic, interacting, nonlinear dynamic processes. Each process is called a unit, and our primary motivation arises in biological metapopulation systems where a unit is a spatially distinct sub-population. Block particle filters are an effective tool for simulation-based likelihood evaluation for these systems, which are strongly dependent through time on a single unit and relatively weakly coupled between units. Iterated filtering algorithms can facilitate likelihood maximization for simulation-based filters.

We introduce a new iterated block particle filter algorithm applicable to parameters that are either unit-specific or shared between units. We demonstrate this algorithm to carry out inference on a coupled epidemiological model for spatiotemporal COVID-19 case report data in 373 cities.

IP15.04 Panel data analysis via mechanistic models

Your name

Carles Breto

Abstract

Panel or longitudinal analysis of dynamic systems using scientifically motivated mechanistic models is becoming increasingly common. Key factors are the increasing availability of data and advances in statistical inference methodology that does not rely on linearity or Gaussian assumptions. Examples of such inference tools are iterated filtering algorithms. When implementation of these algorithms relies on sequential Monte Carlo, care must be taken to make it scalable to large panels. In this talk, I will review the use of iterated filtering algorithms with panel models and illustrate how they can be used to disentangle within- and between-individual features, to deal with parameter weak identifiability and bias thanks to partial pooling of data, and how they can be applied to study the dynamics of infectious diseases.

IP20: "The Interplay Between Causal Inference and Statistical Learning"

13:30 - 15:10 Thursday, 4th January, 2024
EEBrown - Brown Room, Electrical Engineering Building
Chairs Jiwei Zhao
Organiser: Jiwei Zhao

Weining Shen: Causal inference in soccer game analysis

Pan Zhao: A Semiparametric Instrumented Difference-in-Differences Approach to Policy Learning

Yumou Qiu: Inference for Treatment Effects on Many Derived Random Variables

Jiwei Zhao: Assumption Matters: A Semiparametric Analysis of the Average Treatment Effect on the Treated

---

IP20.01 Causal inference in soccer game analysis

Your name

Weining Shen

Abstract

In many sports, it is commonly believed that the home team has an advantage over the visiting team, known as the home field advantage. Yet its causal effect on team performance is largely unknown. This talk introduces a novel causal inference approach to study the causal effect of home field advantage in English Premier League. We develop a hierarchical causal model and show that both league level and team level causal effects are identifiable and can be conveniently estimated. We further develop an inference procedure for the proposed estimators and demonstrate its excellent numerical performance via simulation studies. We implement our method on the 2020-21 English Premier League data and assess the causal effect of home advantage on eleven summary statistics that measure the offensive and defensive performance and referee bias. We find that the home field advantage resides more heavily in offensive statistics than it does in defensive or referee statistics. We also find evidence that teams that had lower rankings retain a higher home field advantage.

IP20.02 A Semiparametric Instrumented Difference-in-Differences Approach to Policy Learning

Your name

Pan Zhao

Abstract

Recently, there has been a surge in methodological development for the difference-in-differences (DiD) approach to evaluate causal effects. Standard methods in the literature rely on the parallel trends assumption to identify the average treatment effect on the treated. However, the parallel trends assumption may be violated in the presence of unmeasured confounding, and the average treatment effect on the treated may not be useful in learning a treatment assignment policy for the entire population. In this article, we propose a general instrumented DiD approach for learning the optimal treatment policy. Specifically, we establish identification results using a binary instrumental variable (IV) when the parallel trends assumption fails to hold. Additionally, we construct a Wald estimator, novel inverse probability weighting (IPW) estimators, and a class of semiparametric efficient and multiply robust estimators, with theoretical guarantees on consistency and asymptotic normality, even when relying on flexible machine learning algorithms for nuisance parameters estimation. Furthermore, we extend the instrumented DiD to the panel data setting. We evaluate our methods in extensive simulations and a real data application.

IP20.03 Inference for Treatment Effects on Many Derived Random Variables

Your name

Yumou Qiu

Abstract

In many applications, the interest is in treatment effects on random quantities of subjects, where those random quantities are not directly observable but can be estimated based on data from each subject. In this paper, we propose a general framework for studying causal inference of this type of problems under a hierarchical data generation setting. The identifiability of causal parameters of interest is shown under a condition on the biasedness of subject level estimates and an ignorability condition on the treatment assignment. Estimation of the treatment effects is constructed by inverse propensity score weighting on the estimated subject level parameters.

A multiple testing procedure able to control the false discovery proportion is proposed to identify the nonzero treatment effects. Theoretical results are developed to investigate the proposed procedure, and numerical simulations are carried out to evaluate its empirical performance. A case study of medication effects on brain functional connectivity of patients with Autism spectrum disorder (ASD) using fMRI data is conducted to demonstrate the utility of the proposed method.

IP20.04 Assumption Matters: A Semiparametric Analysis of the Average Treatment Effect on the Treated

Your name

Jiwei Zhao

Abstract

In this paper, we consider estimation of average treatment effect on the treated (ATT), an interpretable and relevant causal estimand to policy makers when treatment assignment is endogenous. By considering shadow variables that are unrelated to the treatment assignment but related to interested outcomes, we establish identification of the ATT. Then we focus on efficient estimation of the ATT by characterizing the geometric structure of the likelihood, deriving the semiparametric efficiency bound for ATT estimation and proposing an estimator that can achieve this bound. We rigorously establish the theoretical results of the proposed estimator. The finite sample performance of the proposed estimator is studied through comprehensive simulation studies as well as an application to our motivating study.

CP02: Contributed Paper Session

13:30 - 15:10 Thursday, 4th January, 2024
BG03 - Room G03, Babel Building
Chairs Paul Kabaila

Jian Wang: A Bayesian Hierarchical Monitoring Design for Single-Arm Phase II Cancer Clinical Trials

Jing Ning: Enhancing Model Building and Estimating Method Selection: The Crucial Role of Conditional Independence Testing

Mingxuan Cai: Cross-population fine-mapping by leveraging genetic diversity and accounting for confounding bias

Paul Kabaila: The two nested linear regressions testbed for assessing frequentist model averaged confidence intervals

Christopher Baker:

GLM for partially pooled categorical predictors with a case study in biosecurity

CP02.01 A Bayesian Hierarchical Monitoring Design for Single-Arm Phase II Cancer Clinical Trials

Your name

Jian Wang

Abstract

We propose a Bayesian hierarchical monitoring design for single-arm phase II cancer clinical trials that incorporates the information on the duration of response (DOR). To screen a new treatment by evaluating its preliminary therapeutic effect, futility monitoring rules are commonly used in phase II cancer clinical trials to make “go/no-go” decisions timely and efficiently. These futility monitoring rules are usually focused on a single outcome (eg, response rate), although a single outcome may not adequately determine the efficacy of the experimental treatment. To address this issue, we propose Bayesian hierarchical futility monitoring rules to consider both the response rate and DOR. The first level of monitoring evaluates whether the response rate provides evidence that the experimental treatment is worthy of further evaluation. If the evidence from the response rate does not support continuing the trial, the second level monitoring rule, based on the DOR, will be triggered. If both stopping rules are satisfied, the trial will be stopped for futility.

We conducted simulation studies to evaluate the operating characteristics of the proposed monitoring rules and compared them to those of standard method; and illustrated the proposed design with a single arm phase II cancer clinical trial for a combined treatment in patients with relapsed/refractory acute myeloid leukemia. The proposed design avoids an aggressive early termination for futility when the experimental treatment substantially prolongs the duration of response but fails to improve the response rate.

CP02.02 Enhancing Model Building and Estimating Method Selection: The Crucial Role of Conditional Independence Testing

Your name

Jing Ning

Abstract

In the realm of model building and various applications like survival analysis, genetic association studies, and graphical models, determining conditional independence is a challenging but important task. In this presentation, we delve into a novel class of tests for conditional independence, unencumbered by restrictions on conditioning variable distributions. Leveraging the power of generalized Kendall's tau alongside generalized odds ratios, we establish the test statistic's asymptotic properties and present a user-friendly method for acquiring its distribution. One compelling application in assessing the conditional independence assumption is to evaluate the relationship between truncation and failure times given the covariate information. This often-overlooked aspect holds significance for ensuring accurate inference on failure times and determines appropriate estimating methods for the regression analysis on the failure times. Confronting challenges posed by left truncation and right censoring, we introduce tests for this assumption that ingeniously blend the generalized odds ratio from a Cox proportional hazards model with the concept of Kendall's tau. Notably, our method refrains from introducing additional model assumptions beyond the Cox proportional hazards model, while retaining the flexibility of unspecified distributions for truncation times and conditioning variables.

CP02.03 Cross-population fine-mapping by leveraging genetic diversity and accounting for confounding bias

Your name

Cai Mingxuan

Abstract

Fine-mapping prioritizes risk variants identified by genome-wide association studies (GWASs), serving as a critical step to uncover biological mechanisms underlying complex traits. However, several major challenges still remain for existing fine-mapping methods. First, the strong linkage disequilibrium among variants can limit the statistical power and resolution of fine-mapping. Second, it is computationally expensive to simultaneously search for multiple causal variants. Third, the confounding bias hidden in GWAS summary statistics can produce spurious signals. To address these challenges, we develop a statistical method for cross-population fine-mapping (XMAP) by leveraging genetic diversity and accounting for confounding bias. By using cross-population GWAS summary statistics from global biobanks and genomic consortia, we show that XMAP can achieve greater statistical power, better control of false positive rate, and substantially higher computational efficiency for identifying multiple causal signals, compared to existing methods. Importantly, we show that the output of XMAP can be integrated with single-cell datasets, which greatly improves the interpretation of putative causal variants in their cellular context at single-cell resolution.

CP02.04 The two nested linear regressions testbed for assessing frequentist model averaged confidence intervals

Your name

Dr. Paul Kabaila

Abstract

Frequentist model averaged confidence intervals have been proposed by Buckland, Burnham & Augustin, 1997, Biometrics; Fletcher & Turek, 2011, JABA; Turek & Fletcher, 2012, Computational Statistics & Data Analysis. Similar confidence intervals were proposed by Efron, 2014, JASA . There are no comprehensive theoretical results on the coverage and expected length performances of these confidence intervals.

Simulation studies that average performance over randomly-chosen parameter values do not provide stringent assessments.

We use a testbed consisting of two nested linear regressions to provide detailed, stringent and comprehensive performance assessments, in terms of coverage and expected length, of each of these confidence intervals. Any frequentist model averaged confidence interval that performs poorly in this testbed cannot be generally recommended. If a frequentist model averaged confidence interval performs well in this testbed then its performance should be assessed for three nested linear regressions, and so on. Performance results, using the two nested linear regressions testbed, are reported in the following papers.

CP02.05 GLM for partially pooled categorical predictors with a case study in biosecurity

Your name

Christopher Baker

Abstract

National governments use border information to efficiently manage the biosecurity risk presented by travel and commerce. In the Australian border biosecurity system, data about cargo entries are collected from records of directions: that is, the records of actions taken by the biosecurity regulator. An entry is a collection of import lines where each line is a single type of item or commodity. Analysis is simple when the data are recorded in line mode: the directions are recorded individually for each line. The challenge comes when data are recorded in container mode, because the same direction is recorded against each line in the entry, meaning that we don?t know which line(s) within the entry are non-compliant. We develop a statistical model to use container mode data to help inform biosecurity risk of items. We use asymptotic analysis to estimate the value of container mode data compared to line mode data, do a simulation study to verify that we can accurately estimate parameters in a large dataset, and we apply our methods to a real dataset, for which important information about the risk of non-compliance is recovered using the new model.

DL05: Distinguished Lecture Session

15:30 - 17:10 Thursday, 4th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Judith Rousseau

Kerrie Mengersen: Dealing with Sensitive Data

Brodie Lawson: Accessible Fisher-Rao Distances and Fréchet Means for the Multivariate Normals

David Warne: Efficient parameter inference for expensive stochastic models using approximations and transformations

DL05.01 Dealing with Sensitive Data

Your name

Kerrie Mengersen

Abstract

Many datasets of interest to statisticians are subject to privacy conditions. This can constrain access, analysis, sharing and release of results. In this presentation, we will consider two ways in which this issue might be addressed. The first is through federated learning, in which the analysis is undertaken in such a way that the data remain in situ and private. The second is synthetic generation of the data, such that the simulated data retains salient characteristics but retains the required privacy. We provide some extensions to the class of models that can be considered in federated learning, and an overview of synthetic generation of tabular data. The exposition of these ideas will be motivated by the creation of an Australian Cancer Atlas.

This research is in collaboration with QUT colleagues Conor Hassan and Dr Robert Salomone, and is funded by the Australian Research Council and Cancer Council Queensland.

Key reading:

C Hassan, R Salomone, K Mengersen (2023) Federated variational inference methods for structured latent variable models. arXiv preprint arXiv:2302.03314

C Hassan, R Salomone, K Mengersen (2023) Deep generative models, synthetic tabular data and differential privacy: an overview and synthesis. arXiv preprint arXiv:2307.15424

DL05.02 Accessible Fisher-Rao Distances and Fréchet Means for the Multivariate Normals

Your name

Brodie Lawson

Abstract

Information geometry presents a unique, geometric means of understanding and working with families of probability distributions, in which information warps parameter space in a fashion similar to how gravity warps spacetime. In particular, the Fisher metric provides us with a highly compelling measure of statistical distance (the Fisher-Rao distance), which is information-aware (unlike the Wasserstein distance) and behaves like a proper distance (unlike say, the Kullback-Leibler divergence). In this talk, I will use the ubiquitous normal distribution as a familiar distribution family with which to demonstrate these benefits, but also the challenges of working with Riemannian geometry. While avoiding as much differential geometry as possible, this talk will cover intuitive ways of understanding the induced geometry, and approximations and numerical methods for finding Fisher-Rao distances in the multivariate case. Finally, we will consider how these concepts extend to Fréchet means as defined by the Fisher metric, and compare these means to other distributional means in the space of multivariate normal distributions.

DL05.03 Efficient parameter inference for expensive stochastic models using approximations and transformations

Your name

David Warne

Abstract

Stochastic processes are used to model complex systems in almost all fields of science. For scientific inquiry, model calibration and statistical inference is performed to validate theory, estimate parameters, and make predictions. However, in practice statistical inference for stochastic models is challenging since the likelihood function is intractable in many realistic settings. Furthermore, standard likelihood-free or simulation-based inference methods are not feasible when realisations of the stochastic model are computationally expensive. As a result, it is common practice to apply model approximations despite the possibility of substantial bias in the resulting statistical analysis.

In this talk, I will present an overview of recent work that aims to exploit the computational benefit of approximations while correcting for the bias using transformations. Specifically, I will discuss moment-matching transforms, Bayesian score calibration, and generalised likelihood profiles. Using practical examples inspired by inferences tasks in biology, I will demonstrate how these methods enable computationally efficient and accurate inference.

IP28: "Some recent developments in duration analysis with covariates"

15:30 - 17:10 Thursday, 4th January, 2024
B106 - Room 106, Babel Building
Chairs Yingwei Paul Peng
Organiser: Yingwei Paul Peng/Ross Maller

Ross Maller: The Use of Mixture Cure Model Methodology in Medical Applications

Yingwei Paul Peng: Nonparametric cure models

Alice Richardson: The cure rate model in pharmaceutical regulation

Shu-Kay Angus Ng: Joint frailty modelling of time-to-event data with recurrent and terminal events

---

IP28.01 The Use of Mixture Cure Model Methodology in Medical Applications

Your name

Ross Maller

Abstract

The mixture cure model is appropriate in survival analysis when the Kaplan-Meier empirical survival curve of a data set levels off or “plateaus” at its right hand end because the largest or a number of the largest lifetimes are censored. This may indicate the presence of a proportion of individuals who are cured of the disease. We present a brief overview of some statistical methodology developed to deal with this kind of data, stressing aspects and problems which are peculiar to it. In particular, when testing for the presence of cured individuals in the population, we need to have sufficient follow-up in the trial. Additionally, we point out that the assumptions underlying the Cox proportional hazard model are unlikely to be satisfied in this kind of data, and when applied without modification, may be misleading. An important aspect for diagnostic and prognostic purposes is to calculate the probability that an individual, having survived till a designated time, is cured. We illustrate doing this for a classical set of breast cancer data.

IP28.02 Nonparametric cure models

Your name

Yingwei Peng

Abstract

Nonparametric estimation methods for the cure rate and the distribution of the failure time of uncured subjects with covariates for right-censored survival data have attracted much attention in the last few years. To model the effects of covariates on the latency distribution of the failure time of uncured subjects, existing works assume that the cure rate is either constant or depends on the same covariate as the covariate in the latency distribution of uncured subjects. In this talk, I will review the nonparametric estimation methods for the mixture cure model and present a new nonparametric method to model covariate effects on the latency distribution of uncured subjects. The estimation method is based on the EM algorithm, which is readily available for mixture cure models, and it relaxes the assumption used in the existing works. The finite sample and asymptotic properties of the proposed estimator are discussed. Finally, the nonparametric estimation methods are employed to model the effects of some covariates on the time to bankruptcy among commercial banks insured by the FDIC during the first quarter of 2006. This is joint work with Ana López-Cheda and María Amalia Jácome.

IP28.03 The cure rate model in pharmaceutical regulation

Your name

Alice Richardson

Abstract

The cure rate model in survival analysis provides a rich class of models for the apparent levelling of survival curves observed in many experiments and clinical trials involving time to event, in particular progression-free survival times for new treatments for cancer. Yet the models appear underutilised in the important next step for drug development, namely the approval of drugs for use or funding in a particular country. In Australia, these schemes are managed by the Pharmaceutical Benefits Advisory Committee, who advise the Department of Health and thence the Minister for Health on the approval decision.

The first part of this presentation will be a scoping review of the choice of model in recently published results of cancer drug trials. This will be followed by a discussion of the attitudes of pharmaceutical companies and health regulators towards new developments in modelling, and a proposal for a way forward to increase acceptance of new modelling techniques. This part of the presentation will be illustrated with a re-analysis of published data from a recent trial to compare modelling outcomes.

IP28.04 Joint frailty modelling of time-to-event data with recurrent and terminal events

Your name

Shu-Kay Angus Ng

Abstract

We present an innovative perspective on analysing time-to-event recurrent data to elicit the evolution pathway of events of interest. The proposed methodology is based on a joint frailty modelling approach via a generalised linear mixed model (GLMM) formulation to account for the heterogeneous risk of failure and the presence of informative censoring due to a terminal event. The GLMM approach considers two correlated random effects to jointly model the dependence between the hazard rates of recurrent events and the terminal event. Efficient estimation of model parameters is achieved using an extended best linear unbiased prediction (BLUP) and approximate residual maximum likelihood (REML) procedures. The proposed model provides information on event-free survival and mean residual life functions as well as prediction of patient-specific frailties for both recurrent and terminal events. We will demonstrate the capacity of our method using a cancer registry data set of melanoma patients in Australia, regarding the occurrence of secondary primary cancer after the diagnosis of melanoma. We will also present the relative performance of the proposed joint frailty model to a standard frailty model via simulation studies.

Keywords: Joint frailty models, Random effects, Time-to-event data, Mean residual life, Cancer registry data.

IP41: "Recent advances in statistical classification, ranking, and change-point detection"

15:30 - 17:10 Thursday, 4th January, 2024
B305 - Room 305, Babel Building
Chairs Tiejun Tong
Organiser: Tiejun Tong

Lucy Xia: Non-splitting Neyman-Pearson Classifiers EMPTY

Weichen Wang: Spectral Ranking Inferences Based on General Multiway Comparisons

Le Zhou: Sparse Convoluted Rank Regression in High Dimensions

Heng Peng: Automatic Change Point Detection and SegmentEstimation via Variational Bayesian Model Selection

IP41.01 **Non-splitting Neyman-Pearson Classifiers**

Your name

Lucy Xia

Abstract

The Neyman-Pearson (NP) binary classification paradigm constrains the more severe type of error (e.g., the type I error) under a preferred level while minimizing the other (e.g., the type II error). This paradigm is suitable for applications such as severe disease diagnosis, and fraud detection, among others. A series of NP classifiers have been developed to guarantee the type I error control with high probability. However, these existing classifiers involve a sample splitting step: a mixture of class 0 and class 1 observations to construct a scoring function and some left-out class 0 observations to construct a threshold. This splitting enables classifier construction built upon independence, but it amounts to insufficient use of data for training and a potentially higher type II error. Leveraging a canonical linear discriminant analysis (LDA) model, we derive a quantitative CLT for a certain functional of quadratic forms of the inverse of sample and population covariance matrices, and based on this result, develop for the first time NP classifiers without splitting the training sample. Numerical experiments have confirmed the advantages of our new non-splitting parametric strategy.

IP41.02 Spectral Ranking Inferences Based on General Multiway Comparisons

Your name

Weichen Wang

Abstract

This paper studies the performance of the spectral method in the estimation and uncertainty quantification of the unobserved preference scores of compared entities in a very general and more realistic setup in which the comparison graph consists of hyper-edges of possible heterogeneous sizes and the number of comparisons can be as low as one for a given hyper-edge. Such a setting is pervasive in real applications, circumventing the need to specify the graph randomness and the restrictive homogeneous sampling assumption imposed in the commonly-used Bradley-Terry-Luce (BTL) or Plackett-Luce (PL) models. Furthermore, in the scenarios when the BTL or PL models are appropriate, we unravel the relationship between the spectral estimator and the Maximum Likelihood Estimator (MLE). We discover that a two-step spectral method, where we apply the optimal weighting estimated from the equal weighting vanilla spectral method, can achieve the same asymptotic efficiency as the MLE. Given the asymptotic distributions of the estimated preference scores, we also introduce a comprehensive framework to carry out both one-sample and two-sample ranking inferences, applicable to both fixed and random graph settings. It is noteworthy that it is the first time effective two-sample rank testing methods are proposed. Finally, we substantiate our findings via comprehensive numerical simulations and subsequently apply our developed methodologies to perform statistical inferences on statistics journals and movie rankings.

IP41.03 Sparse Convoluted Rank Regression in High Dimensions

Your name

Le Zhou

Abstract

High-dimensional penalized rank regression was studied in 2020, and it was shown to enjoy nice theoretical properties. Compared with the least squares, rank regression can have a substantial gain in estimation efficiency while maintaining a minimal relative efficiency of 86.4%. However, the computation of penalized rank regression can be very challenging for high-dimensional data, due to the highly nonsmooth rank regression loss.

In this work we view the rank regression loss as a non-smooth empirical counterpart of a population level quantity, and a smooth empirical counterpart is derived by substituting a kernel density estimator for the true distribution in the expectation calculation. This view leads to the convoluted rank regression loss and consequently the sparse penalized convoluted rank regression (CRR) for high-dimensional data. We prove some interesting asymptotic properties of CRR. Under the same key assumptions for sparse rank regression, we establish the rate of convergence of the ℓ_1-penalized CRR for a tuning free penalization parameter and prove the strong oracle property of the folded concave penalized CRR. We further propose a high-dimensional Bayesian information criterion for selecting the penalization parameter in folded concave penalized CRR and prove its selection consistency. We derive an efficient algorithm for solving sparse convoluted rank regression that scales well with high dimensions. Numerical examples demonstrate the promising performance of the sparse convoluted rank regression over the sparse rank regression. Our theoretical and numerical results suggest that sparse convoluted rank regression enjoys the best of both sparse least squares regression and sparse rank regression.

IP41.04 Automatic Change Point Detection and SegmentEstimation via Variational Bayesian Model Selection

Your name

Peng Heng

Abstract

Change-point detection has long been an active research area, especially in the Big Data era, where data streams are usually non-stationary. However, many existing methods require a preset number of change points or use a certain stopping threshold, therefore limited to some model selection criterion. In this talk, we introduce an offline Bayesian change point model and associated scalable variational EM algorithm that can automatically estimate the number of change points within parameters in each segment. The comprehensive simulations for a normal mean-variance shift model, a discrete Poisson model, and an actual application in finance demonstrate the advantages of our approach. All the change-point locations and posterior inferences indicate that the proposed method is comparable in location, parameter estimation, and computational efficiency over existing methods.

IP23: "Recent advances in Bayesian computation for complex models"

15:30 - 17:10 Thursday, 4th January, 2024
EEBrown - Brown Room, Electrical Engineering Building
Chairs Nadja Klein
Organiser: Nadja Klein/Matias Quiroz

Nadja Klein: Bayesian function selection in additive models with an application to time-to-event data

Matias Quiroz: A correlated pseudo-marginal approach to doubly intractable problems

David Frazier: Reliable Bayesian Inference in Approximate and Misspecified Models

Robert Kohn: The Block-Correlated Pseudo Marginal Sampler for State Space Models

IP23.01 Bayesian function selection in additive models with an application to time-to-event data

Your name

Nadja Klein

Abstract

Accurately selecting and estimating smooth functional effects in additive models with potentially many functions is a challenging task.We introduce a novel Demmler-Reinsch basis expansion to model the functional effects that allows us to orthogonally decompose an effect into its linear and nonlinear parts, and which has a crucial avantage over the commonly employed mixed model representation of penalized splines. Equipping the reparameterized regression coefficients with normal beta prime spike and slab priors allows us to determine whether a continuous covariate has a linear, a nonlinear or no effect at all. We provide new theoretical results for the prior and provide a compelling explanation for its superior Markov chain Monte Carlo mixing performance compared to alternative methods. We illustrate our approach along effect selection on the hazard rate of a time-to-event response in the geoadditive Cox regression model in simulations and data on survival after leukemia.

IP23.02 A correlated pseudo-marginal approach to doubly intractable problems

Your name

Matias Quiroz

Abstract

Doubly intractable models are encountered in a number of fields, e.g. social networks, ecology and epidemiology. Inference for such models requires the evaluation of a likelihood function, whose normalising function depends on the model parameters and is typically computationally intractable. We propose a signed pseudo-marginal Metropolis-Hastings (PMMH) algorithm with an unbiased block-Poisson estimator to sample from the posterior distribution of doubly intractable models. The advantages of our estimator over previous approaches are that its form is ideal for correlated pseudo-marginal methods which are well known to dramatically increase sampling efficiency. Moreover, we develop analytically derived heuristic guidelines for optimally tuning the hyperparameters of the estimator.

IP23.03 Reliable Bayesian Inference in Approximate and Misspecified Models

Your name

David Frazier

Abstract

We provide a simple and general solution to the fundamental open problem of inaccurate uncertainty quantification of Bayesian inference in misspecified or approximate models, and of generalized Bayesian posteriors more generally. While existing solutions are based on explicit Gaussian posterior approximations, or computationally onerous post-processing procedures, we demonstrate that correct uncertainty quantification can be achieved by substituting the usual posterior with an intuitively appealing alternative posterior conveying the same information. This solution applies to both likelihood-based and loss-based posteriors, and we formally demonstrate the reliable uncertainty quantification of our proposed approach. The new approach is demonstrated through a range of examples, including linear models, and doubly intractable models.

IP23.04 The Block-Correlated Pseudo Marginal Sampler for State Space Models

Your name

Robert Kohn

Abstract

Particle Marginal Metropolis-Hastings (PMMH) is a general approach to Bayesian inference when the likelihood is intractable, but can be estimated unbiasedly. The talk outlines and motivates a PMMH method that scales up better to higher dimensional state vectors than previous approaches, which is achieved by proposing a novel block version of the PMMH and using the trimmed mean of the unbiased likelihood estimates of multiple particle filters; developing an efficient auxiliary disturbance particle filter, which is necessary when the bootstrap disturbance filter is inefficient, but the state transition density cannot be expressed in closed form. The performance of the sampler is investigated empirically by applying it to non-linear Dynamic Stochastic General Equilibrium models and to multivariate GARCH diffusion-driven volatility in the mean models. Although the focus is on applying the method to state space models, the approach will be useful in a wide range of applications such as large panel data models and stochastic differential equation models with mixed effects.

IP47: "Markov chains and related topics"

15:30 - 17:10 Thursday, 4th January, 2024
AW153 - Room 153, Arts West North Building
Chairs Michael Choi
Organiser: Michael Choi

Jun Yang: Stereographic Barker's MCMC Proposal: Efficiency and Robustness at Your Disposal

Geoffrey Wolfer: Improved Estimation of Relaxation Time in Non-reversible Markov Chains

Sumeetpal Singh: On the forgetting of particle filters

Michael Choi: Markov chain entropy games and the geometry of their Nash equilibria

IP47.01 Stereographic Barker's MCMC Proposal: Efficiency and Robustness at Your Disposal

Your name

Jun Yang

Abstract

We introduce a new family of robust gradient-based MCMC samplers under the framework of Stereographic MCMC (Yang et al., 2022) which maps the original high dimensional problem in Euclidean space onto a sphere. Compared with the existing Stereographic Projection Sampler (SPS) which is of a random-walk Metropolis type algorithm, our new family of samplers is gradient-based using the Barker proposal (Livingstone and Zanella, 2022), which improves SPS in high dimensions and is robust to tuning. Meanwhile, the proposed algorithms enjoy all the good properties of SPS, such as uniform ergodicity for a large class of heavy and light-tailed distributions and ``blessings of dimensionality'.

IP47.02 Improved Estimation of Relaxation Time in Non-reversible Markov Chains

Your name

Geoffrey Wolfer

Abstract

The pseudo-spectral gap of a non-reversible ergodic Markov chain, introduced by Paulin [2015], is an important parameter measuring the asymptotic rate of convergence to stationarity. We characterize up to logarithmic factors the minimax trajectory length for the problem of estimating the pseudo-spectral gap of an ergodic Markov chain in constant multiplicative error. Our result recovers the known complexity in the reversible setting for estimating the absolute spectral gap [Levin, Peres 2016, Hsu et al., 2019], and nearly resolves the problem in the general, non-reversible setting. We further improve the known procedure to construct fully empirical confidence intervals for the pseudo-spectral gap. This is joint work with Aryeh Kontorovich (Ben-Gurion University of the Negev).

IP47.03 On the forgetting of particle filters

Your name

Sumeetpal S. Singh

Abstract

We study the forgetting properties of the particle filter when its state --- the collection of particles --- is regarded as a Markov chain. Under a strong mixing assumption on the particle filter's underlying Feynman--Kac model, we find that the particle filter is exponentially mixing, and forgets its initial state in $O(\log N)$ `time', where N is the number of particles and time refers to the number of particle filter algorithm steps, each comprising a selection (or resampling) and mutation (or prediction) operation. An example suggests that this rate is optimal. In contrast to our result, available results to-date are extremely conservative, suggesting $O(\alpha^N)$ time steps are needed, for some $\alpha>1$, for the particle filter to forget its initialisation. We also study the {\it conditional} particle filter (CPF) and extend our forgetting result to this context. We establish a similar conclusion, namely, CPF is exponentially mixing and forgets its initial state in $O(\log N)$ time. To support this analysis, we establish new time-uniform L^p error estimates for CPF, which can be of independent interest.

Joint work with: Anthony Lee (Bristol), Joona Karjalainen (Jyvaskayla) & Matti Vihola (Jyvaskyla).

IP47.04 Markov chain entropy games and the geometry of their Nash equilibria

Your name

Michael Choi

Abstract

Consider the following two-person mixed strategy game of a probabilist against Nature with respect to the parameters (f,B,π), where f is a convex function satisfying certain regularity conditions, B is either the set {L_i}_{i=1}^{n} or its convex hull with each L_i being a Markov infinitesimal generator on a finite state space X and π is a given positive discrete distribution on X. The probabilist chooses a prior measure μ within the set of probability measures on B denoted by P(B) and picks a L∈B at random according to μ, whereas Nature follows a pure strategy to select M∈L(π), the set of ππ-reversible Markov generators on X. Nature pays an amount D_f(M||L), the f-divergence from L to M, to the probabilist. We prove that a mixed strategy Nash equilibrium always exists, and establish a minimax result on the expected payoff of the game. This also contrasts with the pure strategy version of the game where we show a Nash equilibrium may not exist. To find approximately a mixed strategy Nash equilibrium, we propose and develop a simple projected subgradient algorithm that provably converges with a rate of O(1/\sqrt{t}), where t is the number of iterations. In addition, we elucidate the relationships of Nash equilibrium with other seemingly disparate notions such as weighted information centroid, Chebyshev center and Bayes risk. This is based on a joint work with Geoffrey Wolfer (RIKEN AIP), and the paper can be found in https://arxiv.org/abs/2310.04115

IP36:"Recent Advances in Statistical Network Analysis with Applications"

15:30 - 17:10 Thursday, 4th January, 2024
OA239 - Room 239, Old Arts Building
Organiser: Ji Zhu

Yang Feng: Robust Unsupervised Multi-task and Transfer Learning on Gaussian Mixture Models

Junhui Wang: Adaptive Merging and Efficient Estimation in Longitudinal Networks

Jean Yang: Identification of network biomarkers with cross-platform omics prediction

Jinming Li: Statistical Inference on Latent Space Models for Network Data

IP36.01 Robust Unsupervised Multi-task and Transfer Learning on Gaussian Mixture Models

Your name

Yang Feng

Abstract

Unsupervised learning has been widely used in many real-world applications. One of the simplest and most important unsupervised learning models is the Gaussian mixture model (GMM). In this work, we study the multi-task learning problem on GMMs, which aims to leverage potentially similar GMM parameter structures among tasks to obtain improved learning performance compared to single-task learning. We propose a multi-task GMM learning procedure based on the EM algorithm that not only can effectively utilize unknown similarity between related tasks but is also robust against a fraction of outlier tasks from arbitrary distributions. The proposed procedure is shown to achieve minimax optimal rate of convergence for both parameter estimation error and the excess mis-clustering error, in a wide range of regimes. Moreover, we generalize our approach to tackle the problem of transfer learning for GMMs, where similar theoretical results are derived. Finally, we demonstrate the effectiveness of our methods through simulations and real data examples. To the best of our knowledge, this is the first work studying multi-task and transfer learning on GMMs with theoretical guarantees.

IP36.02 A stochastic block Ising model for multi-layer networks with inter-layer dependence

Your name

Jingnan Zhang

Abstract

Community detection has attracted tremendous interests in network analysis, which aims at finding group of nodes with similar characteristics. Various detection methods have been developed to detect homogeneous communities in multi-layer networks, where inter-layer dependence is a widely acknowledged but severely under-investigated issue. In this paper, we propose a novel stochastic block Ising model (SBIM) to incorporate the inter-layer dependence to help with community detection in multi-layer networks. The community structure is modeled by the stochastic block model (SBM) and the inter-layer dependence is incorporated via the popular Ising model. Furthermore, we develop an efficient variational EM algorithm to tackle the resultant optimization task and establish the asymptotic consistency of the proposed method. Extensive simulated examples and a real example on gene co-expression multi-layer network data are also provided to demonstrate the advantage of the proposed method.

IP36.03 Identification of network biomarkers with cross-platform omics prediction

Your name

Jean Yee Hwa Yang

Abstract

In this modern era of precision medicine, molecular signatures identified from advanced omics technologies hold great promise to better guide clinical decisions. However, current approaches are often location-specific due to the inherent differences between platforms and across multiple centres, thus limiting the transferability of molecular signatures. Here we use a Cross-Platform Omics Prediction (CPOP) procedure that accounts for differences in feature scaling in omics data by performing weighted feature selection and estimation that preferentially selects stable features across multiple datasets - thus generating network biomarkers. Next, we conduct an extensive evaluation using melanoma expression data, showing that this biomarker outperforms traditional prediction models based on gene-based features. This demonstrates its potential as an important component in clinical screening for precision medicine. Additionally, we demonstrate the generalizability of this network biomarker across a range of diseases and omics platforms, underscoring its wide applicability.

IP36.04 Statistical Inference on Latent Space Models for Network Data

Your name

Jinming Li

Abstract

Latent space models are powerful statistical tools for modeling and understanding network data. While the importance of accounting for uncertainty in network analysis is well recognized, current literature predominantly focuses on point estimation and prediction, leaving the statistical inference of latent space models an open question. This work aims to fill this gap by providing a general framework to analyze the theoretical properties of the maximum likelihood estimators. In particular, we establish the individual and joint asymptotic distribution results for the latent space models under different edge types and link functions. Furthermore, the proposed framework enables us to generalize our results to the dependent-edge and sparse scenarios. Our theories are supported by simulation studies and have the potential to be applied in downstream inferences, such as link prediction and network testing problems.

CP11: Contributed Paper Session

15:30 - 17:10 Thursday, 4th January, 2024
OA107 - Room 107, Old Arts Building
Chairs Yoshihiko Maesono

Haoze Hou: Non-parametric Estimation of General Heterogeneous Causal Effects with Covariate Measurement Error

Yoshihiko Maesono: Asymptotic properties of kernel type quantile estimators

Gan Yuan: Efficient Estimation of the Central Mean Subspace via Smoothed Gradient Outer Products

Dong Luo: Semi-Supervised Estimation of Marginal Means: An Optimal Constrained Least Squares Approach

Sangkon Oh: Robust mixture of regressions with nonparametric symmetric errors

CP11.01 Non-parametric Estimation of General Heterogeneous Causal Effects with Covariate Measurement Error

Your name

Haoze Hou

Abstract

This paper considers a generalized optimization framework for estimation of general heterogeneous treatment (GHTE) effects when the covariates are exposed to classical measurement errors. The framework includes the conditional average, quantile, and asymmetric least squares causal effects of treatment as special cases. Under the unconfoundedness condition, we show that GHTE can be identified through a weighted optimization based on which we propose deconvolution kernel estimators. We derive the asymptotic bias and variance of our proposed estimators and provide their asymptotic linear expansions, which is useful for the statistical inference in practice. We adopt the simulation-extrapolation method to select the smoothing parameters and propose a new extrapolation procedure to stabilize the computation. Monte Carlo simulations and a real data analysis support the benefits of our estimators under measurement error.

CP11.02 Asymptotic properties of kernel type quantile estimators

Your name

Yoshihiko Maesono

Abstract

In this talk, we will discuss asymptotic properties of the two kernel type quantile estimators. The estimators are proposed from different point of views. Many researchers study quantile estimation because of its application to practical problems in the financial industry and risk assessment. Based on the kernel estimation method, there are two type estimators. One is weighted combination of order statistics and the other is a direct inversion of the kernel distribution function estimator, which is equal to the quantile regression without covariates. In this talk we show that those estimators are asymptotically equivalent, and we compare asymptotic mean squared errors of them.

CP11.03 Nonparametric curve estimation in measurement error problems with conditionally heteroscedastic variances

Your name

Jiyang Zhang

Abstract

We consider the problem of estimating the density fX of a latent variable X using replicated observations contaminated by additive errors depending on X. Standard Fourier techniques for deconvolution seem inappropriate because of this lack of independence, and we suggest instead a basis expansion technique. We show that by choosing a basis of Legendre polynomials, we can express the unknown coefficients of the target density in terms of invertible equations of moments of the contaminated data. Our proposed nonparametric estimator of fX attains optimal minimax convergence rates in the conditional Gaussian error case. To implement our density estimator in practice, we introduce fully data-driven approaches for choosing a truncating parameter and computing an upper bound to the support of fX. We illustrate the numerical performance of our estimator on simulated examples and on replicated data from the National Health and Nutrition Examination Survey.

CP11.04 Semi-Supervised Estimation of Marginal Means: An Optimal Constrained Least Squares Approach

Your name

Dong Luo

Abstract

We develop a new semi-supervised method for estimating the marginal mean of a response variable in scenarios where traditional linear models fall short, while generalized linear models (GLMs) such as logistic or Poisson regression can serve as good working models—even if they are misspecified. Our estimator is handy for handling binary or strictly positive response variables. While the approach involves fitting a GLM on the labelled data and then supplementing with the unlabelled data, we show that the optimal asymptotic variance for the estimator of the marginal mean is obtained via a constrained least squares approach to the GLM fit, rather than via the conventional maximum likelihood. We demonstrate that our estimator is consistent and obtains smaller variance than the sample mean and the use of maximum likelihood whenever there is a non-zero correlation between the response and some covariate, regardless of the correct specification of the GLM form. We provide simulation studies to support our theoretical findings.

CP11.05 Robust mixture of regressions with nonparametric symmetric errors

Your name

Sangkon Oh

Abstract

In finite mixture of regression models, normal assumption for the errors of each regression component is typically adopted. Though this common assumption is theoretically and computationally convenient, it often produces inefficient and undesirable estimates which undermine the applicability of the model particularly in the presence of outliers.

To mitigate these limitations, we propose to use nonparametric Gaussian scale mixture distributions for component error distributions. By this means, we can lessen the risk of misspecification and obtain robust estimators, as the class of Gaussian scale mixture distributions encompasses many essential symmetric unimodal distributions.

In this paper, we study the identifiability of the proposed model and develop a feasible estimating algorithm. Numerical studies including simulation studies and real data analysis to demonstrate the performance of the proposed method are also presented.

CP15: Contributed Paper Session

15:30 - 17:10 Thursday, 4th January, 2024
BG03 - Room G03, Babel Building
Chairs Rami Tabri

Rami Tabri: The Information Projection in Moment Inequality Models: Existence, Dual Representation, and Approximation

Stephen Muirhead: Persistence of stationary Gaussian fields with spectral singularity

Ayesha Perera: Flexible functions: sinc versus cubic spline interpolation

Yicheng Zeng: Signal plus noise models in the log proportional regime: When does debiasing help?

Paul Mostert: Generalized compounding models using recursive quadrature approximation in a Bayesian setting

CP15.01 The Information Projection in Moment Inequality Models: Existence, Dual Representation, and Approximation.

Your name

Rami Tabri

Abstract

The Information projection (I-projection) is a cross-entropy minimizer that has played a significant role in the information theoretic approach to Statistics. It arises in many areas such as Constrained Statistical Inference, Variational Bayesian Inference, Computational Optimal Transport, and Large Deviations Theory.

We present new existence, dual representation, and approximation results for the I-projection in the infinite-dimensional setting for moment inequality constraint sets. These results are established under a general specification of the moment inequalities, nesting both conditional and unconditional moments and allowing for an infinite number of such inequalities. Constraints of this kind are connected to important applications; for example, with imposing shape restrictions on densities and regressions.

An essential innovation of this research is the exhibition of the dual variable as a weak vector-valued integral, enabling the formulation of an approximation scheme of the I-projection's equivalent Fenchel dual problem. In particular, we show under suitable assumptions that the values of finite-dimensional programs can approximate the dual problem's optimum value and that, in addition, every accumulation point of a sequence of optimal solutions for the approximating programs is an optimal solution of the dual problem. We illustrate the verification of assumptions and the construction of the approximation scheme's parameters for the cases of unconditional and conditional first-order stochastic dominance constraints and dominance conditions that characterize selectionable distributions for a random set. Furthermore, we include numerical experiments based on these examples to demonstrate the simplicity of the approximation scheme in practice and its straightforward implementation using off-the-shelf optimization methods.

CP15.02 Persistence of stationary Gaussian fields with spectral singularity

Your name

Stephen Muirhead

Abstract

We consider the probability that a stationary Gaussian field stays positive on a large ball $B(T)$. Our main result is that, assuming the field has a spectral singularity at the origin of order $\alpha \in (0,d)$ (as well as certain mild regularity conditions), this probability decays at log-asymptotic rate $m (d-\alpha)(\log T) \capa(B(T))$, where $\capa(D)$ is the capacity of D with respect to the field, and m is the mass of the absolutely continuous component of its spectrum. If the covariance kernel K is regularly varying at infinity with index $-\alpha$, this rate can be expressed as $c_{d;\alpha} m (d-\alpha)(\log T) K(T)^{-1}$ where $c_{d;\alpha} > 0$ is an explicit constant. This generalises a result of Dembo and Mukherjee on positively-correlated Gaussian processes to a wide class of Gaussian fields with spectral singularity. In particular we do not assume positive correlations. Joint work with Naomi Feldheim and Ohad Feldheim.

CP15.03 Flexible functions: sinc versus cubic spline interpolation

Your name

Ayesha Perera

Abstract

Kabaila & Giri (2009) and Mainzer & Kabaila (2019) considered a linear regression model with normally distributed random errors and parameter of interest a specified linear combination of the regression parameters. They constructed a confidence interval for this parameter that utilizes the uncertain prior information that a distinct linear combination of the regression parameters takes a specified value. This confidence interval is obtained by numerical nonlinear constrained optimization of a confidence interval specified by two functions *b* and *s*. The function *b* is a real-valued odd continuous function and the function *s* is a positive real-valued even continuous function.

These functions were originally specified using cubic spline interpolation with given equally-spaced knots. While these cubic splines are very smooth between successive knots (they are cubics), they have discontinuous third derivatives at the knots. This has a negative impact the computation of integrals that involve the functions *b* and s and need to be evaluated during the numerical optimization. We have chosen, instead, to specify the functions *b* and *s* using sinc function interpolation, resulting in infinitely differentiable functions. This then greatly eases the computation of the integrals that involve the functions b and s and need to be evaluated during the numerical optimization.

CP15.04 Signal plus noise models in the log proportional regime: When does debiasing help?

Your name

Yicheng Zeng

Abstract

We consider the popular signal plus noise model with a low-rank signal and heterogenous noise in the log proportional regime where the sample size and data dimension are comparable in logarithm. In this work, we aim to quantify the effect of the noise on recovering the low-rank signal, and find out the situations where debiasing singular values could improve the statistical accuracy, and also understand the mechanism behind it.

We first derive an explicit form for the bias in the leading singular values of the noise-corrupted data matrix and then add a debiasing step into the signal recovery procedure. Under an entrywise loss measuring the recovery error, we show that this new estimation has a higher convergence rate to zero than the classical hard thresholding estimation when the aspect ratio, i.e., the ratio of dimension and sample size, is divergent. Thus, we conclude from this interesting phenomenon that the debiasing procedure on the singular values could sometimes help us significantly improve the statistical accuracy although the singular vectors remain unchanged, which is a new finding in the case of divergent aspect ratio that is common in the log proportional regime. Furthermore, we use updated concentration inequalities to the local laws from random matrix theory and then derive finite sampled, i.e., non-asymptotic, results for the singular values and vectors, as well as the estimation error. Lastly, we study matrix denoising, multi-dimensional scaling, and clustering as applications.

CP15.05 Generalized compounding models using recursive quadrature approximation in a Bayesian setting

Your name

Paul J Mostert

Abstract

Generalization through compounding increases the versatility and flexibility of assumed statistical models allowing for single parametric models to have multiple parameters controlling their form and use. The additional parameters are responsible for the heavier tails in compound distributions. Paramount to objective Bayesian inference is the fact that the Bernstein-von Mises theorem states that given enough data, the posterior distribution becomes no longer dependent on the prior as such, but rather on the Fisher information. Due to the complexity of the resulted generalized distributions through compounding models, a sub-group of non-informative priors depend exclusively on the Fisher information that in turn not always result in closed-form solutions. Numerical integration or quadrature are some of the more direct approaches to approximate for example the Fisher information when using a class of non-informative priors, while others like the traditional resampling schemes may result in approximations with analytical difficulties. The approach followed in this paper to approximate Fisher information is an adaptive Gauss-Kronrod quadrature with extrapolation by a recursive Wynn's ϵ-algorithm. A simulation study illustrates the effectiveness of Bayesian inference using specific non-informative priors in the presence of some symmetric and asymmetric loss functions within generalized compounding models. An example in daily wind-speed in a suburban area measured over time is used as illustration with these parameter-rich models in a Bayesian setting.

DL06: Distinguished Lecture Session

08:30 - 10:10 Friday, 5th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Aurore Delaigle

Annie Qu: Individualized Dynamic Model for Multi-resolutional Data with Application of Mobile Health

Heping Zhang: Use of mobile data to examine compliance in clinical trials

Peter Song: Learning digital features from wearable device data to advance mobile health

DL06.01 Individualized Dynamic Model for Multi-resolutional Data with Application of Mobile Health

Your name

Annie Qu

Abstract

Mobile health has emerged as a major success in tracking individual health status, due to the popularity and power of smartphones and wearable devices. This has also brought great challenges in handling heterogeneous, multi-resolution data which arise ubiquitously in mobile health due to irregular multivariate measurements collected from individuals. In this talk, we propose an individualized dynamic latent factor model for irregular multi-resolution time series data to interpolate unsampled measurements of time series with low resolution. One major advantage of the proposed method is the capability to integrate multiple irregular time series and multiple subjects by mapping the multi-resolution data to the latent space. In addition, the proposed individualized dynamic latent factor model is applicable to capturing heterogeneous longitudinal information through individualized dynamic latent factors. In theory, we provide the interpolation error bound of the proposed estimator and derive the convergence rate with non-parametric approximation methods. Both the simulation studies and the application to smartwatch data demonstrate the superior performance of the proposed method compared to existing methods.

DL06.02 Use of mobile data to examine compliance in clinical trials

Your name

Heping Zhang

Abstract

Compliance to interventions is very important for behavioral modification trials, but is generally difficult to monitor. We used data from a multi-center trial, called "Improving Reproductive Fitness through Pretreatment with Lifestyle Modification in Obese Women with Unexplained Infertility". We used the data from 358 participants 57,505 observations from this trial. We chose the daily total steps to define a measure of compliance. The trial was composed of multiple sequential phases, so we defined phase-specific compliance scores, and weighed the phase-specific scores to obtain the overall compliance score for each participant. We examined how the compliance scores were associated with other variables, and the association between pregnancy outcomes with the overall compliance score, while adjusting selected covariates. The overall compliance score was significantly associated with baseline variables including baseline steps, education level, history of prior conception, history of prior pregnancy loss, and baseline estradiol level, but not significantly associated with pregnancy outcomes.

DL06.03 Learning digital features from wearable device data to advance mobile health

Your name

Peter Song

Abstract

Wearable devices are becoming a popular mini-robot to collect real-time information of well-being from a device user. Our primary goal is to extract and validate features relevant to personal health conditions from massive high-frequency measurements, so that each device user can utilize personal digital features to guide for personal health management. In this talk, I will introduce several projects that collect wearable devices data to develop cost-effective mobile healthcare programs. Through these diverse public health research projects we hope to demonstrate the use of such data collected from mobile devices to address some important health-related challenges such as monitoring of infectious disease, sleep health, aging, and reproductive health. I will give an overview on the application of different supervised learning techniques to process wearable devices data and build adaptive AI-systems to assist people on health-related decision making.

IP50: "Advances in the development and application of Bayesian nonparametrics models"

08:30 - 10:10 Friday, 5th January, 2024
B106 - Room 106, Babel Building
Chairs Jan Hannig
Organiser: Yuan Ji

Peter Mueller: Constructing synthetic control arms from real world data using nonparametric Bayesian common atoms models

Michele Guindani: Semi-parametric local variable selection under misspecification

Yang Ni: Graphical Dirichlet Process for Clustering Non-Exchangeable Grouped Data

Yuan Ji: PAM: Plaid Atoms Model for Bayesian Nonparametric Analysis of Grouped Data

IP50.01 Constructing synthetic control arms from real world data using nonparametric Bayesian common atoms models

Your name

Peter Mueller

Abstract

We propose a model-based approach using nonparametric Bayesian common atoms models to create synthetic matching populations from available data. The model and method are developed in the context of single-arm, treatment only clinical trials. The single-arm cohort is complemented by a synthetic control arm that is created from readily available external data in the form of electronic health records (EHR). Although randomized clinical trials (RCT) remain the gold standard for approvals by regulatory agencies, the increasing availability of such real world data has opened opportunities to supplement increasingly expensive and difficult to carry out RCT's with evidence from readily available real world data.

We propose a novel nonparametric Bayesian common atoms mixture model that allows us to find equivalent population strata in the EHR and the treatment arm and then resample the EHR data to create equivalent patient populations under both, the single arm trial and the resampled EHR. Resampling is implemented as a density-free importance sampling scheme. Using the synthetic control arm inference for the treatment effect can then be carried out using any method available for RCTs. Alternatively the proposed nonparametric Bayesian model allows straightforward model-based inference. In simulation experiments, the proposed method vastly outperforms alternative methods. We apply the method to supplement single arm treatment-only glioblastoma studies with a synthetic control arm based on historical trials.

IP50.02 Semi-parametric local variable selection under misspecification

Your name

Michele Guindani

Abstract

Local variable selection aims to discover localized effects by assessing the impact of covariates on outcomes within specific regions defined by other covariates. The challenges of local variable selection are outlined in the presence of non-linear relationships and model misspecification. Specifically, a potential drawback of commonly used semi-parametric methods is highlighted: even slight model misspecification can result in a high rate of false positives. To address these shortcomings, a methodology based on orthogonal splines is proposed that achieves consistent local variable selection in high-dimensional scenarios. The approach offers simplicity, handles continuous and discrete covariates, accommodates multivariate covariates, and provides theory for high-dimensional covariates and model misspecification. Settings with either independent or dependent data are discussed. The proposed approach allows for the inclusion of adjustment covariates, enhancing flexibility in modeling complex scenarios. Simulation studies illustrate its application with independent and correlated data and two real datasets. One dataset evaluates salary gaps associated with discrimination factors at different ages, while the other examines the effects of covariates on brain activation over time.

IP50.03 Graphical Dirichlet Process for Clustering Non-Exchangeable Grouped Data

Your name

Yang Ni

Abstract

We consider the problem of clustering grouped data with possibly non-exchangeable groups whose dependencies can be characterized by a known directed acyclic graph. To allow the sharing of clusters among the non-exchangeable groups, we propose a Bayesian nonparametric approach, termed graphical Dirichlet process, that jointly models the dependent group-specific random measures by assuming each random measure to be distributed as a Dirichlet process whose concentration parameter and base probability measure depend on those of its parent groups. The resulting joint stochastic process respects the Markov property of the directed acyclic graph that links the groups.

We characterize the graphical Dirichlet process using a novel hypergraph representation as well as the stick-breaking representation, the restaurant-type representation, and the representation as a limit of a finite mixture model. We develop an efficient posterior inference algorithm and illustrate our model with simulations and a real grouped single-cell dataset.

IP50.04 PAM: Plaid Atoms Model for Bayesian Nonparametric Analysis of Grouped Data

Your name

Yuan Ji

Abstract

We propose a new class of Bayesian nonparametrics models for grouped data called the Plaid Atoms Model (PAM). PAM belongs to a class of dependent random distributions built upon the Dirichlet process. It induces dependent clusters across multiple groups and allows some clusters to be either shared across groups or uniquely possessed by a group. We discuss the construction of PAM via an atom-skip mechanism and discuss theoretical properties of the proposed new models. Minor extensions of the proposed model for multivariate or count data are presented. Simulation studies and applications using real-world datasets illustrate the model's desirable performance.

IP64: "Statistical methods for dimension reduction and high-dimensional regression"

08:30 - 10:10 Friday, 5th January, 2024
OA239 - Room 239, Old Arts Building
Chairs Hong Zhang
Organiser: Hong Zhang

Jing Zeng: Robust Sliced Inverse Regression: Estimation and Optimality for Heavy-Tailed Data in High Dimensions

Liujun Chen: Dimension Reduction for Extreme Tail Regression via Contour Projection

Baihua He: Transfer Learning by Optimal Model Averaging for Censored Data

Zhanfeng Wang: Estimation and model selection for nonparametric function-on-function regression

IP64.01 Robust Sliced Inverse Regression: Estimation and Optimality for Heavy-Tailed Data in High Dimensions

Your name

Jing Zeng

Abstract

Sliced inverse regression (SIR) is a flexible modeling tool that effectively reduces dimensions to reveal complicated mechanism behind data. In recent years, SIR has been generalized to high dimensions in a variety of ways. However, all existing methods rely on the light-tailed assumption for predictors, which is frequently violated in real life. To tackle ubiquitous heavy-tailed data, we propose a novel robust SIR method, referred to as ROSE, that scales well to high dimensions and heavy-tailedness simultaneously. We start with an adaptive distribution model that explicitly incorporates heavy tails, and covers many popular distribution assumptions as special cases. Then ROSE leverages a new elegant invariance result to convert the original SIR problem to a less challenging one on a latent set of light-tailed predictors. We rigorously show that ROSE admits the same minimax optimal convergence rate as existing light-tailed methods even when we only have finite second moments. ROSE is also computationally efficiency compared to existing robust methods in that no extra tuning parameter selection is required for overcoming the heavy-tailedness. Extensive simulation studies and two real data examples are conducted to support the theoretical results.

IP64.02 Dimension Reduction for Extreme Tail Regression via Contour Projection

Your name

Liujun Chen

Abstract

In a variety of regression problems, the central interest is to infer the extremes of the response given a set of predictors. However, the high dimensionality of the predictor usually poses the great challenge to the inference problem. Dimension reduction is an elegant technique to solve the dimensionality issue. In this paper, we introduce the notion of the central extreme subspace (CES), whose existence and uniqueness are guaranteed under the fairly mild condition. After projecting the predictor onto a contour, we develop three target subspaces which find their counterparts in sufficient dimension reduction. The corresponding sample estimators are also developed. Such contour projection procedure assists to connect the target subspaces to the CES. Under mild assumptions, the resulting estimators are shown to achieve favorable finite-sample properties. Moreover, we show that the proposed method is robust in the sense that the finite-sample efficiency is maintained when the predictors exhibit the heavy-tailedness, a common characteristic of many real life data. Our proposal contributes a useful addition to the toolkit of the extreme regression and also greatly expands the realm of dimension reduction. We further demonstrate the effectiveness and robustness of our proposal through the extensive simulation study and the application on two real examples, a concrete data and a financial data.

IP64.03 Transfer Learning by Optimal Model Averaging for Censored Data

Your name

Baihua He

Abstract

Transfer learning has gained significant attention in various domains, addressing the challenge of limited individual study data for prediction. In this paper, we develop a transfer learning approach with model averaging to predict censored responses in the main model. Specifically, several helper models are formulated with shared parameters from other datasets, and the optimal weights for the averaging procedure are derived by minimizing a delete-one cross-validation criterion.

The proposed transfer learning allows the model forms to vary among helper models. We show that the proposed approach achieves the lowest prediction risk asymptotically when the main model is misspecified and attains model weight consistency when the main model is correctly specified. We further demonstrate that the risk of the proposed approach is no larger than the risks of the equal weighting approach and any single candidate model asymptotically, regardless of the correctness of the main model. The performances of the proposed procedure are illustrated and compared with other existing methods on extensive simulation studies and the Surveillance, Epidemiology, and End Results (SEER)-Medicare liver cancer data.

IP64.04 **Estimation and model selection for nonparametric function-on-function regression**

Your name

Zhanfeng Wang

Abstract

Regression models with a functional response and functional covariate have received significant attention recently. While various nonparametric and semiparametric models have been developed, there is an urgent need for model selection and diagnostic methods. In this article, we develop a unified framework for estimation and model selection in nonparametric function-on-function regression. We propose a general nonparametric functional regression model with the model space constructed through smoothing spline analysis of variance (SS ANOVA). The proposed model reduces to some of the existing models when selected components in the SS ANOVA decomposition are eliminated. We propose new estimation procedures under either L1 or L2 penalty and show that the combination of the SS ANOVA decomposition and L1 penalty provides powerful tools for model selection and diagnostics. We establish consistency and convergence rates for estimates of the regression function and each component in its decomposition under both the L1 and L2 penalties. Simulation studies and real examples show that the proposed methods perform well.

IP18: "Recent Progress in Statistics Analysis of Random Field Models"

08:30 - 10:10 Friday, 5th January, 2024
OA107 - Room 107, Old Arts Building
Chairs Yimin Xiao
Organiser: Yimin Xiao

David Bolin: Statistical modeling of data on metric graphs

Cheuk Yin Lee: Strong local nondeterminism for a class of fractional SPDEs

Yimin Xiao: Some Recent Results on Multivariate Gaussian Random Fields

Saifei Sun: Fixed-domain asymptotics for Gaussian random fields

IP18.01 Statistical modeling of data on metric graphs

Your name

David Bolin

Abstract

There is a growing interest in the statistical modeling of data on compact metric graphs such as street or river networks based on Gaussian random fields. In this work, we introduce the Whittle-Matérn fields, which is a class of models specified as solutions to a fractional-order stochastic differential equation on the metric graph. Contrary to earlier covariance-based approaches for specifying Gaussian fields on metric graphs, the Whittle-Matérn fields are well-defined for any compact metric graph and can provide Gaussian
processes with differentiable sample paths given that the fractional exponent is large enough.

We present some of the main statistical properties of the models, discuss spatio-temporal extensions, and finally illustrate the usage of the models through an application to traffic data, where we use the recently introduced MetricGraph R package to fit and compare different models.

IP18.02 Strong local nondeterminism for a class of fractional SPDEs

Your name

Cheuk Yin Lee

Abstract

We consider a parametric class of stochastic fractional diffusion equations driven by additive Gaussian noise that is white noise or possibly correlated both in time and space. This generates a rich family of Gaussian random fields. We give necessary and sufficient conditions (Dalang's condition) for the existence of random field solutions. We also prove strong local nondeterminism for the solution as a temporal process or a spatial process. As a result, this allows us to deduce sharp regularity properties for the solutions. This is joint work with Le Chen and Panqiu Xia.

IP18.03 Some Recent Results on Multivariate Gaussian Random Fields

Your name

Yimin Xiao

Abstract

In this talk, we present some recent results on the statistical properties of several classes of multivariate Gaussian random fields including operator fractional Brownian motion. These results illustrate the effects of the dependence structures among the coordinate processes on prediction and sample path properties of multivariate Gaussian random fields.

IP18.04 Fixed-domain asymptotics for Gaussian random fields

Your name

Saifei Sun

Abstract

The project considers the covariance parameter estimation for Gaussian random fields that are observed with measurement error and irregularly spaced design sites on a fixed and bounded domain. The Gaussian random fields are assumed to have smooth mean functions and isotropic covariance functions belonging to powered exponential, Matérn, or generalized Wendland class. Under fixed-domain asymptotics, consistent estimators are proposed for three microergodic parameters, namely the nugget, the smoothness parameter, and a parameter related to the coefficient of the principal irregular term of the covariance function. Upper bounds for the convergence rate of these estimators are also established.

IP29: "Current Topics in Statistical Analysis of High Dimensional Genomic Data"

08:30 - 10:10 Friday, 5th January, 2024
B305 - Room 305, Babel Building
Chairs Somnath Datta

Geoffrey McLachlan: The detection of differentially expressed genes via cluster-specific contrasts of mixed effects

Tapabrata Maiti: Error Controlled Feature Selection for Ultrahigh Dimensional and Highly Correlated Feature Space Using Deep Learning

Subharup Guha: A zero-inflated Bayesian nonparametric approach for identifying differentially abundant taxa in multigroup microbiome data with covariates

Somnath Datta: A pseudo-value regression approach for differential network analysis of co-expression data

IP29.01 The detection of differentially expressed genes via cluster-specific contrasts of mixed effects

Your name

Geoffrey McLachlan

Abstract

The detection of differentially expressed (DE) genes whose expression levels vary between two or more classes representing different experimental conditions (say, diseases), is a commonly studied problem in bioinformatics. We consider further our approach to this problem based on a test statistic formed as a weighted (normalized) cluster-specific contrast in the mixed effects of the mixture model used in the first instance to cluster the gene profiles into a manageable number of clusters. The key factor in the formation of our test statistic is the use of gene-specific mixed effects in the cluster-specific contrast. It thus means that the (soft) assignment of a given gene to a cluster is not crucial. This is because in addition to class differences between the (estimated) fixed effects terms for a cluster, gene-specific class differences also contribute to the cluster-specific contributions to the final form of the test statistic. The proposed test statistic can be used where the primary aim is to rank the genes in order of evidence against the null hypothesis of no DE. We also show how a P-value can be calculated for each gene for use in multiple hypothesis testing where the intent is to control the false discovery rate (FDR) at some desired level.

IP29.02 Error Controlled Feature Selection for Ultrahigh Dimensional and Highly Correlated Feature Space Using Deep Learning

Your name

Taps Maiti

Abstract

Deep learning has been at the center of analytics in recent years due to its impressive empirical success in analyzing complex data objects. Despite this success, most existing tools behave like black-box machines, thus the increasing interest in interpretable, reliable, and robust deep learning models applicable to a broad class of applications. Feature-selected deep learning has emerged as a promising tool in this realm. However, the recent developments do not accommodate ultra-high dimensional and highly correlated features or high noise levels. In this article, we propose a novel screening and cleaning method with the aid of deep learning for a data-adaptive multi-resolutional discovery of highly correlated predictors with a controlled error rate. Extensive empirical evaluations over a wide range of simulated scenarios and several real datasets demonstrate the effectiveness of the proposed method in achieving high power while keeping the false discovery rate at a minimum.

IP29.03 A zero-inflated Bayesian nonparametric approach for identifying differentially abundant taxa in multigroup microbiome data with covariates

Your name

Subharup Guha

Abstract

Scientific studies conducted during the last two decades have established the central role of the microbiome in disease and health. Differential abundance analysis aims to identify microbial taxa associated with two or more sample groups defined by attributes such as disease subtype, geography, or environmental condition. The results, in turn, help clinical practitioners and researchers diagnose disease and develop new treatments more effectively. However, detecting differential abundance is uniquely challenging due to the high dimensionality, collinearity, sparsity, and compositionality of microbiome data. Further, there is a critical need for unified statistical approaches that can directly compare more than two groups and appropriately adjust for covariates.

We develop a zero-inflated Bayesian nonparametric (ZIBNP) methodology that meets the multipronged challenges posed by microbiome data and identifies differentially abundant taxa in two or more groups while accounting for sample-specific covariates. The proposed hierarchical model flexibly adapts to unique data characteristics, casts the typically high proportion of zeros in a censoring framework, and mitigates high dimensionality and collinearity issues by utilizing the dimension-reducing property of the semiparametric Chinese restaurant process. The approach relates the microbiome sampling depths to inferential precision and conforms with the compositional nature of microbiome data. In simulation studies and in the analyses of the CAnine Microbiome during Parasitism (CAMP) dataset on infected and uninfected dogs and the Global Gut microbiome dataset on human subjects belonging to three geographical regions, we compare ZIBNP with established statistical methods for differential abundance analysis in the presence of covariates.

IP29.04 A pseudo-value regression approach for differential network analysis of co-expression data

Your name

Somnath Datta

Abstract

A differential network (DN) analysis identifies changes in measures of association among genes under two or more experimental conditions. In this talk, we introduce a pseudo-value regression approach for network analysis (PRANA). This is a novel method of differential network analysis that also adjusts for additional clinical covariates. We start from mutual information criteria, followed by pseudo-value calculations, which are then entered into a robust regression model. To the best of our knowledge, this is the first attempt of utilizing a regression modeling for DN analysis by collective gene expression levels between two or more groups with the inclusion of additional clinical covariates. We study its performance through simulations, and on a real data from the Gene Expression Omnibus database to identify DC genes that are associated with chronic obstructive pulmonary disease to demonstrate its utility. By and large, adjusting for available covariates improves accuracy of a DN analysis.

IP38: "Statistical theory for deep learning"

08:30 - 10:10 Friday, 5th January, 2024
AW153 - Room 153, Arts West North Building
Chairs Susan Wei
Organiser: Susan Wei

Peter Bartlett: In-Context Learning Linear Models with Transformers

Sophie Langer: Dropout in the Linear Model

Masaaki Imaizumi: Statistical Analysis on Generalization Ability of In-Context Learning

Dino Sejdinovic: Squared Neural Families of Tractable Densities and Intensities

---

IP38.01 In-Context Learning Linear Models with Transformers

Your name

Peter Bartlett

Abstract

Transformer networks have demonstrated a remarkable ability at in-context learning (ICL): given a short prompt sequence of labeled data, they can behave like supervised learning algorithms. We investigate the dynamics of ICL in transformers with a single linear self-attention layer trained by gradient flow on linear regression tasks. We show that despite non-convexity, gradient flow with a suitable initialization finds a global optimum and achieves prediction error competitive with the best linear predictor over the test prompt distribution, but it is not robust to shifts in the covariate distribution. For a simplified parameterization, we establish a statistical complexity bound for attention model pretraining using stochastic gradient descent, showing how in-context learning performance improves with the number of independent tasks.

Based on joint work with Ruiqi Zhang and Spencer Frei, and with Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, and Quanquan Gu.

IP38.02 Dropout in the Linear Model

Your name

Sophie Langer

Abstract

Dropout is a technique to prevent neural networks from overfitting. By randomly dropping neurons in each training step, it can be seen as an algorithmic regularization technique, that has proven effectiveness in various applications. Most theoretical results link dropout with explicit regularization. In case of linear models the relation to l2-penalization was stated. In this talk we investigate the explicit behavior of the iterates generated by gradient descent with dropout in a linear model. We present two different ways of applying dropout in the training procedure. First, by taking the gradient of a model reduced by dropout. Second, by taking the gradient of the full model and then reducing it by dropout. In the first, more complex, case we show that in expectation the iterates converge to a l2-regularized estimator, while the sequence of variances does not converge to its explicit counterpart. In the second case the convergence of the iterates to the linear least squares estimator is shown. Our analysis is explicit, stating convergence rates, that depend on the input, the learning rate, the initalization point and the dropout probability.

This talk is based on joint work with Gabriel Clara and Johannes Schmidt- Hieber.

IP38.03 Statistical Analysis on Generalization Ability of In-Context Learning

Your name

Masaaki Imaizumi

Abstract

In this study, we analyze sample complexity in in-context learning, a type of meta-learning. In-context learning is a framework that consists of an identical learner capable of handling multiple tasks and has attracted strong attention in recent artificial intelligence technologies. As an approach to understanding this learning framework, several studies have raised a hypothesis that the learner learns an algorithm itself. In this study, we study this hypothesis of algorithmic learning through statistical sample complexity analysis.

Specifically, we evaluate the generalization ability of in-context learning using task selection and prompt length, as well as the complexity of the mapping on an empirical distribution. Through these quantitative assessments, we try to gain a better understanding of in-context learning.

IP38.04 Squared Neural Families of Tractable Densities and Intensities

Your name

Dino Sejdinovic

Abstract

Flexible models for probability distributions are an essential ingredient in many machine learning tasks. We develop and investigate a new class of probability distributions, which we call a Squared Neural Family (SNEFY), formed by squaring the 2-norm of a neural network and normalising it with respect to a base measure. Following the reasoning similar to the well established connections between infinitely wide neural networks and Gaussian processes, we show that SNEFYs admit a closed form normalising constants in many cases of interest, thereby resulting in flexible yet fully tractable density models. SNEFYs strictly generalise classical exponential families, are closed under conditioning, and have tractable marginal distributions. Their utility is illustrated on a variety of density estimation and conditional density estimation tasks.

CP14: Contributed Paper Session

08:30 - 10:10 Friday, 5th January, 2024
BG03 - Room G03, Babel Building
Chairs Kanta Naito

Kanta Naito: Simultaneous Confidence Region of an Embedded One-Dimensional Curve in Multi-Dimensional Space

Natsumi Makigusa: Nonlinear support vector regression with penalized likelihood

Hyeok Kyu Kwon: Minimax optimal density estimation using a shallow generative model with a one-dimensional latent variable

Hui Zhang: Unbiased and Robust Analysis of Co-localization in Super-resolution Microscope Images

CP14.01 Simultaneous Confidence Region of an Embedded One-Dimensional Curve in Multi-Dimensional Space

Your name

Kanta Naito

Abstract

We discuss the simultaneous confidence region of a one-dimensional curve embedded in multi-dimensional space. Local linear regression is applied component-wise to each variable in multi-dimensional data, which yields an estimator of the one-dimensional curve. A simultaneous confidence region of the curve is proposed based on this estimator and its gradient vector. The proposed region is a thin region that contains the estimates of the one-dimensional curve; for example, in the case of three-dimensional data, it looks like a sausage-shaped region. Theoretical results for the region are developed under some reasonable assumptions. Practically efficient algorithms to determine the thickness of the region are also addressed. The effectiveness of the region is investigated through simulation studies and applications to artificial and real datasets, which show that the proposed simultaneous confidence region works well.

CP14.02 Nonlinear support vector regression with penalized likelihood

Your name

Natsumi Makigusa

Abstract

Support vector regression is a method that is known to be robust to outliers and to be able to consider an extension to nonlinear version. In support vector regression, an optimization problem is solved by adding a penalty term to a loss function called ε-insensitive loss. Typically, support vector regression involves some hyperparameters, furthermore, nonlinear support vector regression uses some kernel which also includes additional hyperparameters.Therefore implementing nonlinear support vector regression needs to determine several hyperparameters simultaneously. Such determination of hyperparameters is usually done by cross validation, which is a heavy task in nonlinear support vector regression.

To overcome this issue, we introduce a Laplace-type distribution as the conditional distribution of the response given the explanatory variables, which is determined by the hyperparameters involved in nonlinear support vector regression. This makes it possible to determine the hyperparameters by using likelihood. In particular, we propose nonlinear support vector regression with penalized likelihood method.

Theoretical results for the proposed nonlinear support vector regression are reported. The results of Monte Carlo simulation and of applying the proposed method to some real datasets are also reported.

CP14.03 Minimax optimal density estimation using a shallow generative model with a one-dimensional latent variable

Your name

Hyeok Kyu Kwon

Abstract

The deep generative model yields an implicit estimator for the unknown distribution or density function of the observation. This paper investigates some statistical properties of the implicit density estimator pursued by VAE-type methods from a nonparametric density estimation framework.

More specifically, we obtain convergence rates of the VAE-type density estimator under the assumption that the underlying true density function belongs to a locally Hölder class. Remarkably, a near minimax optimal rate with respect to the Hellinger metric can be achieved by the simplest network architecture, a shallow generative model with a one-dimensional latent variable. The proof of the main theorem relies on the well-known result from the nonparametric Bayesian literature that a smooth density with a suitably decaying tail can efficiently be approximated by a finite mixture of normal distributions. We also discuss an alternative proof, which offers important insights and suggests a potential extension to structured density estimation.

CP14.04 Unbiased and Robust Analysis of Co-localization in Super-resolution Microscope Images

Your name

Hui Zhang

Abstract

Spatial data from high-resolution images abound in many scientific disciplines. For example, single-molecule localization microscopy, such as stochastic optical reconstruction microscopy, provides super-resolution images to help scientists investigate co-localization of proteins and hence their interactions inside cells, which are key events in living cells. However, there are few accurate methods for analyzing co-localization in super-resolution images. The current methods and software are prone to produce false-positive errors and are restricted to only 2-dimensional images. In this paper, we develop a novel statistical method to effectively address the problems of unbiased and robust quantification and comparison of protein co-localization for multiple 2- and 3-dimensional image datasets. This method significantly improves the analysis of protein co-localization using super-resolution image data, as shown by its excellent performance in simulation studies and an analysis of light chain 3-lysosomal-associated membrane protein 1 protein co-localization in cell autophagy. Moreover, this method is directly applicable to co-localization analyses in other disciplines, such as diagnostic imaging, epidemiology, environmental science, and ecology.

CP08: Contributed Paper Session

08:30 - 10:10 Friday, 5th January, 2024
OA224 - Room 224, Old Arts Building
Chairs Xuan Liang

Yuto Miyatake: Modelling the discretization error of initial value problems using the Wishart distribution

Xuan Liang: Covariance Regression for Panel Data with Fixed Effects

Fan Wang: Multilayer random dot product graphs: Estimation and online change point detection

Wei Li: Sufficient dimension reduction in the presence of non-ignorable missing covariates

---

CP08.01 Modelling the discretization error of initial value problems using the Wishart distribution

Your name

Yuto Miyatake

Abstract

Traditional approaches to estimating parameters in ordinary differential equations (ODEs) often involve fitting numerical solutions to observed data. These numerical solutions are usually obtained through numerical integration methods like Runge-Kutta. However, these methods often neglect to consider the discretization errors inherent in numerical integration, thereby limiting the accuracy of parameter estimates.

To improve these estimates, it is crucial to quantify the discretization error. Previous efforts have attempted this by modeling the discretization error as random variables. In this talk, we introduce a new approach that employs the Wishart distribution to model this error. We show that our model not only well quantifies the discretization error but also captures the correlation between variables, thereby enhancing the reliability of parameter estimates in ODEs.

CP08.02 Covariance Regression for Panel Data with Fixed Effects

Your name

Xuan Liang

Abstract

We propose a covariance regression model to study the cross-sectional dependence in panel data with fixed effects, encompassing both individual and time effects. It not only links the covariance of responses to either time-invariant or time-varying similarity matrices induced by auxiliary information, but also establishes the relationship between the mean of responses and covariates. The parameters of interest include both the coefficients of the covariates in the mean of responses, and the coefficients of the similarity matrices in the covariance of responses. Under this new model setting, however, the consistency of the covariance estimator based on the standard concentrated likelihood approach for the fixed effects breaks down. Hence, we propose the restricted quasi-maximum likelihood estimation and the constrained least squares estimation methods. We demonstrate that the resulting estimators are consistent and asymptotically normal, and thoroughly investigate their performances via extensive simulations, and an empirical example.

CP08.03 Multilayer random dot product graphs: Estimation and online change point detection

Your name

Fan Wang

Abstract

We study the multilayer random dot product graph (MRDPG) model, an extension of the random dot product graph to multilayer networks. By modelling a multilayer network as an MRDPG, we deploy a tensor-based method and demonstrate its superiority over existing approaches. Moving to dynamic MRDPGs, we focus on online change point detection problems. At every time point, we observe a realisation from an MRDPG. Across layers, we assume shared common node sets and latent positions but allow for different connectivity matrices. We propose efficient algorithms for both fixed and random latent position cases, minimising detection delay while controlling false alarms. Notably, in the random latent position case, we devise a novel nonparametric change point detection algorithm with a kernel estimator in its core, allowing for the case when the density does not exist, accommodating stochastic block models as special cases. The paper is available on https://arxiv.org/abs/2306.15286.

CP08.04 Sufficient dimension reduction in the presence of non-ignorable missing covariates

Your name

Wei Li

Abstract

Inverse regression based sufficient dimension reduction (SDR) is a popular and powerful class of methods aimed at finding a small number of linear combinations of covariates (known as sufficient predictors) by reversing the regression direction, which retain the full regression information between the response and covariates. Although there exists a large literature on inverse regression based dimension reduction when dealing with a sample of complete data, comparably little has been done on how to perform SDR in the presence of missingness in the covariates.

This article proposes an SDR method when the covariates are missing in an non-ignorable manner, by utilizing a model-based inverse regression approach coupled with a latent missingness mechanism incorporating the response and all covariates. As a fully model-based approach to SDR, we are able to utilize maximum likelihood estimation and inference, and do so via an expectation-maximization type algorithm combined with a clever multi-stage initialization procedure, which involves the multivariate skew-normal distribution family based on an alternative model specification.

Numerical studies demonstrate that the our proposed method can perform better than other traditional SDR methods when non-ignorable missingness exists among covariates. We illustrate an application of the proposed approach to perform SDR on diabetes status of Indians with Pima Indian heritage, where some predictors are incompletely observed.

PL02: Jianqing Fan

10:30 - 11:30 Friday, 5th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Aurore Delaigle

PL02 Factor Augmented Sparse Throughput Deep ReLU Neural Networks for High Dimensional Regression

Your name

Jianqing Fan

Abstract

We introduce a Factor Augmented Sparse Throughput (FAST) model that utilizes both latent factors and sparse idiosyncratic components for nonparametric regression. The FAST model bridges factor models on one end and sparse nonparametric models on the other end. It encompasses structured nonparametric models such as factor augmented additive model and sparse low-dimensional nonparametric interaction models and covers the cases where the covariates do not admit factor structures. Via diversified projections as estimation of latent factor space, we employ truncated deep ReLU networks to nonparametric factor regression without regularization and to more general FAST model using nonconvex regularization, resulting in factor augmented regression using neural network (FAR-NN) and FAST-NN estimators respectively. We show that FAR-NN and FAST-NN estimators adapt to unknown low-dimensional structure using hierarchical composition models in nonasymptotic minimax rates. We also study statistical learning for the factor augmented sparse additive model using a more specific neural network architecture. Our results are applicable to the weak dependent cases without factor structures. In proving the main technical result for FAST-NN, we establish new a deep ReLU network approximation result that contributes to the foundation of neural network theory. Our theory and methods are further supported by simulation studies and an application to macroeconomic data. (Joint work with Yihong Gu)

PL03: Bin Yu

11:30 - 12:30 Friday, 5th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Minge Xie

PL03 Seeking Boolean Interactions in Practice and Theory

Your name

Bin Yu

Abstract

Thresholding or Boolean behaviors of biomolecules underlie many biological processes. Decision-trees capture such behaviors and tree-based methods such random forests have been shown to succeed in predictive tasks in genomics and medicine. In this talk, we use UKBiobank data and a stable version of the random forests, iterative random forests (iRF), to recommend gene and gene-gene interactions that have predictive and stable data evidence for possibly driving a heart disease, Hypertrophic Cardiomyopathy (HCM). Gene-silencing experiments show significant causal evidence in 4 out of the 5 experiments based on iRF-based recommendations and domain knowledge. This and other empirical successes of iRF motivate a theoretical investigation of its tractable version under a new local sparse and spiky (LSS) model where the regression function is a linear combination of Boolean interactions of features. The tractable version of iRF is shown to be model selection consistent under this new model and conditions of feature independence and non-overlap of interactions.

If time allows, I willd describe a new and improved importance measure MDI+ based on a new class of methods, RF+ and with an overlay guided by our Predictabilty-Computability-Stability (PCS) framework for veridical data science.

DL10: Distinguished Lecture Session

13:50 - 15:30 Friday, 5th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Aihua Xia

Qiman Shao: Self-Normalized Cramer Type Moderate Deviation Theorem For Gaussian Approximation

Dennis Leung: Berry-Esseen bounds for Studentized U-statistics

Zhuosong Zhang: Berry–Esseen bounds for Generalized U-statistics

---

DL10.01 SELF-NORMALIZED CRAMER TYPE MODERATE DEVIATION THEOREM FOR GAUSSIAN APPROXIMATION

Your name

Qi-Man Shao

Abstract

Berry–Esseen type bounds for Gaussian approximation of standardized sums have been extensively studied under exponential type moment conditions. In this talk, we shall establish a Cramér type moderate deviation theorem for self-normalized Gaussian approximation under finite moment conditions. More specifically, let $X_{1,} X_2, . . . ,X_n$ be i.i.d. R^p-valued random vectors with zero means. Let $S_{nj} = \sum_i X_{ij}$, $V_{nj}^2 = \sum_i X_{ij}^2$. We show that if the correlation of X_1 is I_p and the third moment of X_1 is finite, then $P(\max_{1 \leq j \leq p} S_{nj} / V_{nj} > x) / P(\max_{1 \leq j \leq p} Z_j > x) \to 1$ uniformly for $0 \leq x \leq o(n^{1/6})$ and for all $p \geq 1$, where $(Z_1, . . . ,Z_p)$ is a normal random vector with zero means and the same correlation matrix as X_1. Similar result is also established for large x when X_1 has a general correlation matrix. The proof is based on a new Cramér type moderate deviation theorem for the minimum of several self-normalized sums.

DL10.02 Berry-Esseen bounds for Studentized U-statistics

Your name

Dennis Leung

Abstract

Establishing limit theorems, such as Berry-Esseen (B-E) bounds, for Studentized statistics has always been more challenging than for standardized statistics, but the resulting theory is arguably more relevant to applications, as most statistics have to be "self-normalized" by a data-driven estimate of their own standard deviations in practice.

In this talk, we will discuss our recent uniform and non-uniform B-E bounds established for Studentized U-statistics of any given degree m, which cover Gosset's t-statistic as a special case with m =1. Under the Stein-method paradigm, a central proof device is a refined exponential randomized concentration inequality originating in the work of Shao (2010). Moreover, we highlight our nonuniform bound, since its form is not a direct parallel of the typical non-uniform bound known for the standardized U-statistics; while counterexamples (Novak, 2005) have shown that the non-uniform bound form for standardized U-statistics is invalid for their Studentized counterparts, our result restores its validity by augmenting the bound with an additive factor that decays exponentially in the sample size n.

DL10.03 Berry–Esseen bounds for Generalized U-statistics

Your name

Zhuosong Zhang

Abstract

We establish optimal Berry–Esseen bounds for the generalized U- statistics. The proof is based on a new Berry–Esseen theorem for exchangeable pair approach by Stein's method under a general linearity condition setting. As applications, an optimal convergence rate of the normal approximation for subgraph counts in Erdös–Rényi graphs and graphon-random graph is obtained.

IP03: "Rare events, risk and zero-one laws"

13:50 - 15:30 Friday, 5th January, 2024
OA239 - Room 239, Old Arts Building
Chairs Gennady Samorodnitsky
Organiser: Arijit Chakrabarty

Gennady Samorodnitsky: Do large deviations cluster? If yes, how?

Thomas Mikosch: Ratios of homogeneous functionals acting on a heavy-tailed time series

Moumanti Podder: Bond percolation games and their generalizations on rooted Galton-Watson trees

Arijit Chakrabarty: Inhomogeneous Erdős-Rényi random graphs: bulk and edge of the spectrum

IP03.01 Do large deviations cluster? If yes, how?

Your name

Gennady Samorodnitsky

Abstract

We describe the cluster of large deviations events that arise when one such large deviations event occurs. We work in the framework of an infinite moving average process with a noise that has finite exponential moments. We consider both short memory and long memory regimes, and show that large deviations cluster differently in these situations.

IP03.02 Ratios of homogeneous functionals acting on a heavy-tailed time series

Your name

Thomas Mikosch

Abstract

We present joint limit theory for sums, maxima and l^p-norms of a multivariate heavy-tailed stationary time series. Heavy tails are defined via regular variation of the finite-dimensional distributions.

We use regular variation calculus introduced by Basrak and Segers (2009, SPA) for stationary sequences: the limiting objects are expressed in terms of the spectral tail process.

Consequences of the limit theory are asymptotic results for ratios of the aforementioned quantities, including Greenwood statistics, studentized sums, ratios of sums and maxima. Particular emphasis is given to self-normalized sums of infinite-variance time series when alpha-stable limit laws for normalized sums appear.

IP03.03 Bond percolation games and their generalizations on rooted Galton-Watson trees

Your name

Moumanti Podder

Abstract

Given a rooted Galton-Watson tree T_{χ} with offspring distribution χ, we assign to each edge of T_{χ}, independently, a label that reads \emph{trap} with probabililty p, \emph{target} with probability q, and \emph{safe} with probability $1-p-q$, with $p+q > 0$. A move involves relocating a token from where it is currently located, say a vertex u of T_{χ}, to any child v of u. Two players take turns to make moves, and a player wins if she is able to move the token along an edge labeled a target, or force her opponent to move the token along an edge labeled a trap, or confine her opponent to a leaf vertex. We analyze the probabilities of the various outcomes associated with this game, and the phase transition phenomenon pertaining to the probability of draw. We generalize this to the toll-tax games where each edge (u,v) of T_{χ} is assigned, independently, a random weight $w(u,v)$ with $P[w(u,v) = 0] = p_{0}, P[w(u,v) = 1] = p_{1}, P[w(u,v) = -1] = p_{-1}$. When a player moves the token along an edge (u,v), she is awarded an amount equal to $w(u,v)$. A player wins if she is the first to amass a capital of amount k_{1} or her opponent is the first to have her capital dwindle to amount $-k_{2}$.

IP03.04 Inhomogeneous Erdős-Rényi random graphs: bulk and edge of the spectrum

Your name

Arijit Chakrabarty

Abstract

The talk is on inhomogeneous Erdős-Rényi random graphs in the non-dense regime. The eigenvalues of the adjacency matrix of the graph are studied. The empirical spectral distribution of the matrix after suitable scaling and centering is shown to have a deterministic limit in probability. Depending on the rank of the inhomogeneity kernel generating the random graph, the largest few eigenvalues have a much higher magnitude than that of the bulk. Assuming the rank to be finite, the second order behaviour of those few eigenvalues, after suitable centring and scaling, is shown to be multivariate Gaussian. The asymptotic behaviour of the corresponding eigenvectors is also studied.

The talk is based on joint works with Sukrit Chakraborty, Rajat Hazra, Frank den Hollander and Matteo Sfragara.

IP60: "Advancements in Geostatistical Analysis: From Earth to Space"

13:50 - 15:30 Friday, 5th January, 2024
EEBrown - Brown Room, Electrical Engineering Building
Chairs Hsin-cheng Huang
Organiser: Hsin-Cheng Huang

Noel Cressie: WOMBAT: Bayesian Inference on Carbon-Dioxide Surface Fluxes from Satellite Data

Francisco Juan Mandujano Reyes: Spatio-temporal dynamic modeling of wildlife disease data

Tingjin Chu: A geostatistical analysis of metallicity variations in galaxies

Nan-Jung Hsu: Estimation of Nonstationary Space Deformation Using Affine Coupling

IP60.01 WOMBAT: Bayesian Inference on Carbon-Dioxide Surface Fluxes from Satellite Data

Your name

Noel Cressie (with Michael Bertolacci, and Andrew Zammit-Mangion)

Abstract

Locations across Earth's surface where the leading greenhouse gas, carbon dioxide (CO_2), is added to or removed from the atmosphere, are known as CO_2 sources and sinks. CO_2 flux is the rate at which this happens, and a critical goal of carbon-cycle science is to characterise the pattern and scale of sources and sinks in both space and time. There is considerable variability in CO_2 fluxes: For example, temperate forests occupy large parts of the terrestrial biosphere and transition from sinks to sources during the year, while volcanoes are local sources with sporadic and unpredictable outgassing of CO_2. Human activity has also caused changes to the natural processes that cause these sources and sinks. In this talk, a framework called WOMBAT (WOllongong Methodology for Bayesian Assimilation of Trace-gases) is presented for predicting CO_2 fluxes in space and time; see Zammit-Mangion et al. (2022; *Geoscientific Model Development*) and Bertolacci et al. (2023; *Annals of Applied Statistics*). WOMBAT is fully Bayesian and produces both spatio-temporal predictions and quantifications of their uncertainties. The framework allows scientists and policy makers to take into account uncertainty in CO_2-flux predictions and, consequently, to produce better mitigation/adaptation strategies for climate change.

IP60.02 Spatio-temporal dynamic modeling of wildlife disease data

Your name

Juan Francisco Mandujano Reyes

Abstract

Statistical methods informed by partial differential equations (PDE) and in particular, reaction-diffusion PDEs known as ecological diffusion equations (EDE) have been studied and used to model wildlife disease spatio-temporal processes. This type of models provides a mechanistic description of the diffusion and growth/decline of a dispersing population or pathogen in space and time. However, the large spatial scales of these ecological processes give rise to intensive computations when fitting spatio-temporal data to complex EDE models using classical numerical methods. In this talk, we explore fast-fitting alternatives using the stochastic partial differential equations (SPDE) Bayesian method featured in the R-INLA package. We consider a stochastic extension of the EDE (SEDE) and discuss its interpretation and main differences from the deterministic EDE. We then study non-stationary extensions of the diffusion-based Gaussian Matérn field and show that these extensions enjoy SEDE-like behavior. The elucidated connection enables us to find a finite element approximated solution for SEDEs by means of the computationally efficient SPDE Bayesian method. For illustration, we analyze the evolution of white-nose syndrome (WNS), an infectious fungal disease in bats in North America, comparing three derived models: stationary SEDE, non-stationary SEDE, and a non-stationary pseudo-SEDE. Finally, a simulation study is conducted to assess the deviance information criterion as a way to differentiate data generated by the three models, as well as the identifiability of the model parameters.

IP60.03 A geostatistical analysis of metallicity variations in galaxies

Your name

Tingjin Chu

Abstract

The metallicity of diffuse ionised gas (DIG) cannot be determined using strong emission line diagnostics, which are calibrated to calculate the metallicity of Hii regions. Because of this, resolved metallicity maps from integral field spectroscopy (IFS) data remain largely incomplete.

In this work, we introduce the geostatistical technique of universal kriging, which allows the complete 2D metallicity distribution of a galaxy to be reconstructed from metallicities measured at Hii regions, accounting for spatial correlations between nearby data points. We apply this method to construct metallicity maps of the local spiral galaxy NGC 5236 using data from the TYPHOON/PrISM survey.

IP60.04 Estimation of Nonstationary Space Deformation Using Affine Coupling

Your name

Nan-Jung Hsu

Abstract

For modeling non-stationary spatial processes, spatial deformation is a popular approach in the literature, which characterizes the underlying non-stationary process as a stationary counterpart in the deformed space through proper space mapping. Existing studies commonly follow a two-step procedure, involving space mapping exploration in the initial step and spatial covariance estimation in the subsequent step. In particular, the first step typically involves estimating local variograms from data, followed by applying a multi-dimensional scaling technique with spline fittings to construct a smoothed deformed space. This study introduces a novel approach to space deformation, based on affine coupling for space mapping. The proposed method applies to both single and multiple realizations of spatial data, offering flexibility and ensuring a bijective mapping. For inference, we use maximum likelihood to simultaneously estimate the deformation space mapping and the covariance model. The effectiveness of the proposed method is demonstrated through simulations and real data examples.

IP06: "Recent Developments of Statistical Methods for Microbiome Research"

13:50 - 15:30 Friday, 5th January, 2024
AW153 - Room 153, Arts West North Building
Chairs Tianying Wang
Organiser: Gen Li

Hongzhe Li: Phylogenetic Association Analysis with Conditional Rank Correlation

Kim-Anh Le Cao: Managing batch effects in microbiome data

Lei Liu: A flexible quasi-likelihood model for microbiome abundance count data

Tianying Wang: A Semiparametric Quantile Single-Index Model for Microbiome Sequencing Count Data

IP06.01 Phylogenetic Association Analysis with Conditional Rank Correlation

Your name

Hongzhe Li

Abstract

Phylogenetic association analysis plays a crucial role in investigating the correlation between microbial compositions and specific outcomes of interest in microbiome studies. However, existing methods for testing such associations have limitations related to the assumption of a linear association in high-dimensional settings and the handling of confounding effects. Therefore, there is a need for methods capable of characterizing complex associations, including non-monotonic relationships.

This paper introduces a novel phylogenetic association analysis framework and associated tests to address these challenges by employing conditional rank correlation as a measure of association. These tests account for confounders in a fully nonparametric manner, ensuring robustness against outliers and the ability to detect diverse dependencies. The proposed framework aggregates conditional rank correlations for subtrees using a weighted sum and maximum approach to capture both dense and sparse signals. The significance level of the test statistics is determined by calibrating through a nearest neighbor bootstrapping method, which is straightforward to implement and can accommodate additional data sets when available. The practical advantages of the proposed framework are demonstrated through numerical experiments utilizing both simulated and real microbiome data sets.

IP06.02 Managing batch effects in microbiome data

Your name

Kim-Anh Lê Cao

Abstract

Microbial communities are highly dynamic and sensitive to changes in the environment. Therefore, microbiome data are highly susceptible to batch effects, defined as sources of unwanted variation that are not related to and obscure any factors of interest.

Existing batch effect correction methods have been primarily developed for gene expression data. As such, they do not consider the inherent characteristics of microbiome data, including zero inflation, overdispersion and correlation between variables.

We introduce new multivariate and non-parametric batch effect correction methods based on Partial Least Squares Discriminant Analysis (PLSDA). PLSDA-batch first estimates treatment and batch variation with latent components, then subtracts batch-associated components from the data. The resulting batch-effect-corrected data can then be input in any downstream statistical analysis. Two variants are proposed to handle unbalanced batch x treatment designs and to avoid overfitting when estimating the components via variable selection.

We compare our approaches with popular methods managing batch effects, namely, removeBatchEffect, ComBat and Surrogate Variable Analysis, in simulated and three case studies using various visual and numerical assessments. We show that our three methods lead to competitive performance in removing batch variation while preserving treatment variation, especially for unbalanced batch treatment designs. Our downstream analyses show selections of biologically relevant taxa. This work demonstrates that batch effect correction methods can improve microbiome research outputs.

IP06.03 A flexible quasi-likelihood model for microbiome abundance count data

Your name

Lei Liu

Abstract

In this paper we present a flexible model for microbiome count data. We consider a quasi-likelihood framework, in which we do not make any assumptions on the distribution of the microbiome count except that its variance is an unknown but smooth function of the mean. By comparing our model to the negative binomial generalized linear model (GLM) and Poisson GLM in simulation studies, we show that our flexible quasi-likelihood method yields valid inferential results. Using a real microbiome study, we demonstrate the utility of our method by examining the relationship between adenomas and microbiota. We also provide an R package 'fql' for the application of our method.

IP06.04 A Semiparametric Quantile Single-Index Model for Microbiome Sequencing Count Data

Your name

Tianying Wang

Abstract

We consider the complex data modeling problem motivated by the zero-inflated and overdispersed microbiome read count data. Analyzing how microbiome abundance is associated with human biological features, such as BMI, is of great importance for host health. Methods based on parametric distributional assumptions, such as zero-inflated Poisson and zero-inflated Negative Binomial regression, have been widely used in modeling such data, yet the parametric assumptions are restricted and hard to verify in real-world applications.

We relax the parametric assumptions and propose a semiparametric single-index quantile regression model. It is flexible to include a wide range of possible association functions and adaptable to the various zero proportions across subjects, which overcomes the major challenge for existing quantile single-index models. We establish the asymptotic properties for the index coefficients estimator and quantile regression curve estimation.

Through extensive simulation studies, we demonstrate the superior performance of the proposed method regarding both model fitting and hypothesis testing. In the application of a microbiome study, compared to existing methods, our method identified more taxa associated with various human biological features, which are also supported by other literature.

IP09: "Machine Learning Methods for Electronic Health Record Data Analysis"

13:50 - 15:30 Friday, 5th January, 2024
OA107 - Room 107, Old Arts Building
Chairs Fei Zou
Organiser: Baiming Zou

Rebecca Hubbard: Classical and machine learning imputation approaches to missing data in electronic health records-based studies

Jian Kang: Optimizing Medical Decision-Making with Reinforcement Learning

Lana Garmire: Deep learning-based prognosis models accurately predict the time to delivery among preeclampsia patients using electronic health record

Baiming Zou: A Deep Neural Network Two-part Model and Feature Importance Test for Semi-continuous Data

IP09.01 Classical and machine learning imputation approaches to missing data in electronic health records-based studies

Your name

Rebecca Hubbard

Abstract

Electronic health records (EHR)-derived data represent an enormous research resource with exposures and outcomes for a large and diverse population. However, EHR data have many limitations including complex patterns of missing data induced by the irregularity of interaction between patients and the healthcare system. Novel approaches to handling missing data including machine learning (ML)-based imputation methods have been touted as a potential solution to this problem, but evaluation of their performance in the context of real-world comparative effectiveness research is lacking. I will present a comparative evaluation of classical and ML methods to addressing missing data in the context of a real-world study of the comparative effectiveness of immunotherapy and chemotherapy for treatment of advanced urothelial cancer. Using plasmode simulation grounded in this context, we compare the performance of traditional and ML imputation methods. We identify settings for missing data in which ML approaches have promise and those in which the greater flexibility of these methods potentially results in overfitting and poor statistical performance.

IP09.02 Reinforcement Learning for Medical Decision-Making on Vaccination

Your name

Jian Kang

Abstract

This research explores the application of reinforcement learning (RL) in guiding medical decision-making based on electronic health records. The decision-making process is conceptualized as a Markov Decision Process (MDP), targeting the identification of the optimal policy that maximizes a defined value function. This function underscores rewards that find a balance between the risk of infections, potential side effects of treatments, and the benefits of maintaining a disease-free status. The underlying risk of infection or disease is specified through a generative model, considering both individual attributes and disease-related factors. To validate this approach, simulations were conducted on synthesized datasets, each encompassing numerous subjects with extended records. The simulated outcomes, stratified by individual attributes, were congruent with real-world data. These findings underscore the potential applicability of the proposed RL framework in shaping informed medical decisions, with special relevance to vaccination applications.

IP09.03 Deep learning-based prognosis models accurately predict the time to delivery among preeclampsia patients using electronic health record

Your name

Lana Garmire

Abstract

Preeclampsia (PE) is one of the leading factors in maternal and perinatal mortality and morbidity worldwide. Delivery timing is key to balancing the risk between severe maternal and neonatal morbidities in pregnancies complicated by PE. Towards this, we developed and validated first-of-their-kind deep learning models that can predict the time to delivery of PE patients at initial diagnosis using electronic health records (EHR) data. The discovery cohort consisted of 1,533 PE cases, including 374 early-onset preeclampsia (EOPE), that were delivered at the University of Michigan Health System (UM) between 2015 and 2021. The validation cohort contained 2,172 PE cases (547 EOPE) from the University of Florida Health System (UF) in the same period. Using the neural network algorithm (Cox-nnet), we constructed baseline and full time-to-delivery prediction models for all PE patients and a subset of EOPE patients.

The baseline models used 45 demographic, medical history, comorbidity, PE severity, and initial diagnosis gestational age features, while the full models included an additional 15 lab tests and vital signs features. To enhance interpretability and accuracy, the models underwent re-training and validation with reduced sets of the most crucial features.The 7-feature baseline models on all PE patients reached C-indices of 0·74 and 0·73 on UM hold-out testing and UF validation dataset respectively, whereas the 12-feature full model had improved C-indices of 0·79 and 0·74 on the same datasets. For the EOPE cases, the 6-feature baseline model achieved C-indices of 0·68 and 0·63, and its 13-feature full model counterpart reached C-indices of 0·76 and 0·67 in the same datasets. Besides confirming the prognostic importance of gestational age at the time of diagnosis and of severe PE status, all four models identified parity and PE in prior pregnancies as important features, which are not in the current guidelines for PE delivery timing. We set up a user-friendly web interface to allow personalized PE time to delivery prediction. The app is available at http://garmiregroup.org/PE-prognosis-predictor/app.

IP09.04 A Deep Neural Network Two-part Model and Feature Importance Test for Semi-continuous Data

Your name

Baiming Zou

Abstract

Semi-continuous data frequently arise in clinical practice. For example, while many surgical patients suffer from varying degrees of acute postoperative pain (POP) sometime after surgery, others experience none, indicating the existence of two distinct data processes at play. Existing parametric or semi-parametric two-part modeling methods for this type of semi-continuous data can fail to appropriately model these two underlying data processes as such methods rely heavily on (generalized) linear additive assumptions. However, many factors may interact to jointly influence the experience of POP non-additively and non-linearly. We derive a deep neural networks (DNN)-based two-part model to address this challenge by adapting the conventional DNN methods, an approach we denote as sDNN. To improve the interpretability and transparency of sDNN, we further derive a feature importance testing procedure to identify important features associated with the outcome measurements, denoting this approach fsDNN. We show that fsDNN not only offers a statistical inference procedure for each feature under complex association but also that using the identified features can further improve the predictive performance of sDNN.

The proposed sDNN- and fsDNN-based two-part models are applied to the analysis of real data from a POP study, in which application they clearly demonstrate advantages over the existing parametric and semi-parametric two-part models. Further, we conduct extensive numerical studies and draw comparisons with other machine learning methods to demonstrate that sDNN and fsDNN consistently outperform the existing two-part models and frequently use machine learning methods regardless of the data complexity.

IP07: "Bayesian methods for detecting key features in complex applications"

13:50 - 15:30 Friday, 5th January, 2024
B106 - Room 106, Babel Building
Chairs Subhashis Ghoshal
Organiser: Subhashis Ghoshal

Michael Zhang: Bayesian Non-linear Latent Variable Modeling via Random Fourier Features

Anindya Roy: Bayesian Graph Estimation Under Causal Vector Autoregressive Time Series

Seonghyun Jeong: Synergizing Roughness Penalization and Basis Selection in Bayesian Spline Regression

Subhashis Ghoshal: Optimal Bayesian Smoothing of Functional Observations over a Large Graph

---

IP07.01 Bayesian Non-linear Latent Variable Modeling via Random Fourier Features

Your name

Michael Zhang

Abstract

The Gaussian process latent variable model (GPLVM) is a popular probabilistic method used for nonlinear dimension reduction, matrix factorization, and state-space modeling. Inference for GPLVMs is computationally tractable only when the data likelihood is Gaussian. Moreover, inference for GPLVMs has typically been restricted to obtaining maximum a posteriori point estimates, which can lead to overfitting, or variational approximations, which mischaracterize the posterior uncertainty. Here, we present a method to perform Markov chain Monte Carlo (MCMC) inference for generalized Bayesian nonlinear latent variable modeling. The crucial insight necessary to generalize GPLVMs to arbitrary observation models is that we approximate the kernel function in the Gaussian process mappings with random Fourier features; this allows us to compute the gradient of the posterior in closed form with respect to the latent variables. We show that we can generalize GPLVMs to non-Gaussian observations, such as Poisson, negative binomial, and multinomial distributions, using our random feature latent variable model (RFLVM). Our generalized RFLVMs perform on par with state-of-the-art latent variable models on a wide range of applications, including motion capture, images, and text data for the purpose of estimating the latent structure and imputing the missing data of these complex data sets.

IP07.02 Bayesian Graph Estimation Under Causal Vector Autoregressive Time Series

Your name

Anindya Roy

Abstract

Multivariate time series data are routinely collected in many application areas. Although stationarity, causality and invertibility are very useful modeling assumptions for any time series data, methodological developments are limited under these assumptions for multivariate time series. Under some assumptions on the autocovariance matrices, in this article, we achieve those properties for a new class of Gaussian multivariate time series. In this proposed class, the normalized multivariate time series is assumed to be some orthogonal rotation of a set of independent univariate latent time series. To capture the graphical dependence structure among the variables we also propose to sparsely estimate the marginal precision matrix and develop related computational methodologies. An efficient Markov Chain Monte Carlo algorithm is developed for posterior computation. We also study theoretical consistency properties. Via limited simulation experiments and real data analysis, we show that the proposed method performs well in practice.

IP07.03 Synergizing Roughness Penalization and Basis Selection in Bayesian Spline Regression

Your name

Seonghyun Jeong

Abstract

Bayesian P-splines and basis determination through Bayesian model selection are both commonly employed strategies for nonparametric regression using spline basis expansions within the Bayesian framework. Although both methods are widely employed, they each have particular limitations that may introduce potential estimation bias depending on the nature of the target function. To overcome the limitations associated with each method while capitalizing on their respective strengths, we propose a new prior distribution that integrates the essentials of both approaches. The proposed method assesses the complexity of the spline model based on a penalty term formed by a convex combination of the penalties from both methods.

The proposed method exhibits adaptability to the unknown level of smoothness while achieving the minimax-optimal posterior contraction rate up to a logarithmic factor. We provide an efficient Markov chain Monte Carlo algorithm for implementing the proposed method. Our extensive simulation study and application to a real dataset substantiate the validity of our proposed approach.

IP07.04 Optimal Bayesian Smoothing of Functional Observations over a Large Graph

Your name

Subhashis Ghoshal

Abstract

Occasionally, some types of data are observed in high-resolution, essentially continuously in time. Such data units are best described as taking values in a space of functions. Subject units carrying the observations may have intrinsic relations among themselves, and are best described by the nodes of a large graph. It is often sensible to think that the underlying signals in these functional observations vary smoothly over the graph, in that neighboring nodes have similar underlying signals. This qualitative information allows the borrowing of strength from neighboring nodes and consequently leads to more accurate inference. We consider a model with Gaussian functional observations and adopt a Bayesian approach to smoothing over the nodes of the graph. We characterize the minimax rate of estimation in terms of the regularity of the signals and their variation across nodes, quantified in terms of the graph-Laplacian. We build an appropriate prior from the graph-Laplacian that attains the minimax bound, while a mixture prior can reach the minimax rate up to a logarithmic factor simultaneously for all possible values of functional and graphical smoothness. We also show that in the fixed smoothness setting, an optimally sized credible region has arbitrarily high frequentist coverage. A simulation experiment demonstrates that the method performs better than potential competing methods like the random forest. The method is also applied to a dataset on daily temperatures measured at several weather stations in the US state of North Carolina. This is a joint work with Arkaprava Roy.

IP44: "Advances in Survival Analysis"

13:50 - 15:30 Friday, 5th January, 2024
BG03 - Room G03, Babel Building
Chairs Jun Ma
Organiser: Jun Ma

Aishwarya Bhaskaran: Accelerated failure time models under partly interval censoring and time-varying covariates

Youngjo Lee: On the Statistical Foundations of H-likelihood for Unobserved Random Variables

Il Do Ha: Deep Neural Networks for Semi-parametric Frailty Models

Yunwei Zhang: Evaluation of the impact of left-censoring on the validity of time-dependent propensity score matching method

IP44.01 Accelerated failure time models under partly interval censoring and time-varying covariates.

Your name

Aishwarya Bhaskaran

Abstract

Accelerated failure time (AFT) models are frequently used for modelling survival data. It is an appealing approach as it asserts a direct relationship between the time to event and covariates, wherein the failure times are either accelerated or decelerated by a multiplicative factor in the presence of these covariates. Several methods exist in the current literature for fitting semiparametric AFT models with time-fixed covariates. However, most of these methods do not easily extend to settings involving both time-varying covariates and partly interval censored data. We propose a maximum penalized likelihood approach to fit a semiparametric AFT model with both time-fixed and time-varying covariates, for survival data with partly interval censored failure times.

IP44.02 On the Statistical Foundations of H-likelihood for Unobserved Random Variables

Your name

Youngjo Lee

Abstract

The maximum likelihood estimation is widely used for statistical inferences. In this study, we reformulate the h-likelihood proposed by Lee and Nelder in 1996, whose maximization yields maximum likelihood estimators for fixed parameters and asymptotically best unbiased predictors for random parameters. We establish the statistical foundations for h-likelihood theories, which extend classical likelihood theories to embrace broad classes of statistical models with random parameters. The maximum h-likelihood estimators asymptotically achieve the generalized Cramer-Rao lower bound. Furthermore, we explore asymptotic theory when the consistency of either fixed parameter estimation or random parameter prediction is violated. The introduction of this new h-likelihood framework enables likelihood theories to cover inferences for a much broader class of models, while also providing computationally efficient fitting algorithms to give asymptotically optimal estimators for fixed parameters and predictors for random parameters.

IP44.03 Deep Neural Networks for Semi-parametric Frailty Models

Your name

IL DO HA

Abstract

For prediction of clustered time-to-event data, we propose a new deep neural network-based frailty model (DNN-FM). An advantage of the proposed model is that the joint maximization of the new h-likelihood provides maximum likelihood estimators for fixed parameters and best unbiased predictors for random frailties. Thus, the proposed DNN-FM is trained by using a negative profiled h-likelihood as a loss function, constructed by profiling out the non-parametric baseline hazard. Experimental studies show that the proposed method enhances the prediction performance of the existing methods (e.g. DNN based Cox model) and provides the feature selection using the multi-head attention. A real data analysis shows that the inclusion of subject-specific frailties to the DNN-Cox model helps to improve the risk prediction of the DNN based Cox model.

IP44.04 Evaluation of the impact of left-censoring on the validity of time-dependent propensity score matching method.

Your name

Yunwei Zhang

Abstract

The propensity score (PS) matching approach forms pairs of treated and non-treated subjects with similar score values to study the average treatment effect of the treated individuals. One variation of PS matching, the time-dependent PS matching approach, has been developed to effectively handle time-dependent treatment. In this case, the PS is estimated using survival models such as the Cox model, as opposed to standard logistic regression. In registry-based studies, the disease onset time is often left-censored because subjects joined the registry after diagnosis. When interest lies in quantifying the effect of this disease on another disease, we could consider the use of PS matching. But, the validity of PS matching with left-censored time-dependent treatment variables has not been thoroughly explored.

Motivated by the presence of left-censored diabetes onset time among National Diabetes Services Scheme subjects used to investigate the effect of diabetes on aging-related diseases such as the dementia, our study aims to comprehensively examine the impact of left censoring on the validity of the time-dependent PS.

In this talk, we will first introduce our approach, where we consider 17 different scenarios where left censoring is not random but instead depends on variables that may be associated with both treatment and outcome variables. Within each scenario, two different time-dependent PS approaches are examined. Our results show the validity of the time-dependent PS matching approach depends on the amount of left censoring present. This highlights the need to carefully apply the time-dependent PS when left censoring is present in the data.

IP35: "Recent development in Causal Inference"

13:50 - 15:30 Friday, 5th January, 2024
B305 - Room 305, Babel Building
Chairs Wei Huang
Organiser: Wei Huang

Zheng Zhang: Causal Quantile Regression for A Continuous Treatment with A Diverging Number of Confounders

Mingming Gong: Identifiable causal generative models for out-of-distribution generalization

Shuo Li: Estimation and Inference for Extreme Continuous Treatment Effects

Lin Liu: High-dimensional semiparametric learning under minimal sparsity conditions

IP35.01 Causal Quantile Regression for A Continuous Treatment with A Diverging Number of Confounders

Your name

Zhang Zhang

Abstract

This paper proposes a varying coefficient model for the continuous quantile treatment effects. We estimate the inverse general propensity score function of growingdimensional confounders using artificial neural networks (ANNs). We establish a new approximation error bound for the ANNs to the nuisance function belonging to a mixed Sobolev class without a known sparsity structure. We show that the ANNs can alleviate the “curse of dimensionality” under this circumstance. We establish the consistency and asymptotic normality of the proposed quantile treatment effects estimator, and apply a weighted bootstrap procedure for conducting inference. The proposed methods are illustrated via simulation studies and a real data application.

IP35.02 Identifiable causal generative models for out-of-distribution generalization

Your name

Mingming Gong

Abstract

Fair machine learning aims to avoid treating individuals or sub-populations unfavourably based on sensitive attributes, such as gender and race. Those methods in fair machine learning that are built on causal inference ascertain discrimination and bias through causal effects. Though causality-based fair learning is attracting increasing attention, current methods assume the true causal graph is fully known. This paper proposes a general method to achieve the notion of counterfactual fairness when the true causal graph is unknown. To select features that lead to counterfactual fairness, we derive the conditions and algorithms to identify ancestral relations between variables on a Partially Directed Acyclic Graph (PDAG), specifically, a class of causal DAGs that can be learned from observational data combined with domain knowledge.

IP35.03 Estimation and Inference for Extreme Continuous Treatment Effects

Your name

Shuo Li

Abstract

This paper studies estimation and inference for the treatment effect on deep tails of the potential outcome distributions corresponding to a continuously valued treatment, namely the extreme continuous treatment effect. We consider two measures for the tail characteristics: the quantile function and the tail mean function defined as the conditional mean beyond a quan- tile level. Then for a quantile level close to 1, we define the extreme quantile treatment effect (EQTE) and extreme average treatment effect (EATE), which are respectively the differences of the quantile and tail mean at different treatment statuses. We propose estimators for the EQTE and EATE based on tail approximations from the extreme value theory. Our limiting theory is for the EQTE and EATE processes indexed by a set of quantile levels and hence facilitates uniform inference for the EQTE and EATE over multiple tail levels. Simulations suggest that our method works well in finite samples and two empirical studies illustrate its practical merits.

IP35.04 High-dimensional semiparametric learning under minimal sparsity conditions

Your name

Lin Liu

Abstract

Treatment effect estimation under unconfoundedness is a fundamental task in causal inference. In response to the challenge of analyzing high-dimensional datasets collected in substantive fields such as epidemiology, genetics, economics, and social sciences, many methods for treatment effect estimation with high-dimensional nuisance parameters (the outcome regression and the propensity score) have been developed in recent years. However, it is still unclear what is the necessary and sufficient sparsity condition on the nuisance parameters for the treatment effect to be $\sqrt{n}$-estimable. In this paper, we propose a new Double-Calibration strategy that corrects the estimation bias of the nuisance parameter estimates computed by regularized high-dimensional techniques and demonstrate that the corresponding Doubly-Calibrated estimator achieves $\sqrt{n}$-rate as long as one of the nuisance parameters is sparse with sparsity below $\sqrt{n}/\log p$, where p denotes the ambient dimension of the covariates, whereas the other nuisance parameter can be arbitrarily complex and completely misspecified. The Double-Calibration strategy can also be applied to settings other than treatment effect estimation, e.g. regression coefficient estimation in the presence of a diverging number of controls in a semiparametric partially linear model.

CP03: Contributed Paper Session

13:50 - 15:30 Friday, 5th January, 2024
OA224 - Room 224, Old Arts Building
Chairs Monitirtha Dey

Ye Yuan: Tropospheric formaldehyde levels infer ambient formaldehyde-induced brain diseases and global burden in China, 2013–2019

Jaehee Kim: Dynamic Functional Connectivity Change-Point Detection with Random Matrix Theory Inference for Brain Network Data

Biswadeep Ghosh: Parametric Analysis of Bivariate Current Status data with Competing risks using Frailty models

Monitirtha Dey: On Limiting Behaviors of Stepwise Multiple Testing Procedures

Kanae Takahashi: Statistical inferences for the F1-score of multi-label classification

CP03.01 Tropospheric formaldehyde levels infer ambient formaldehyde-induced brain diseases and global burden in China, 2013–2019

Your name

Ye Yuan

Abstract

Our objective was to examine the association between tropospheric airborne pollutants and human health risk and global burden, especially, attributable to indoor formaldehyde (FA) pollution in China. The data of tropospheric pollutants, such as: CO, NO, O3, PM2.5 or PM10, SO2, and FA in China, 2013–2019, which were derived from the database of satellite remote-sensing, were first calculated and then analyzed them according to satellite cloud pictures. The rate of prevalence, incidence, deaths, years of life lost (YLLs), years lived with disability (YLDs), and disability-adjusted life-years (DALYs) of the Chinese population was obtained from the Global Burden of Diseases (GBD 2010). A linear regression analysis was used to evaluate the relationship between tropospheric FA concentrations and GBD indexes of human brain diseases, the numbers of fire plot, the average summer temperature, population density and car sales in China from 2013 to 2019.

Our results showed that the levels of tropospheric FA could reflect the degree of indoor air FA pollution on a nationwide scale in China; in particular, only tropospheric FA exhibited a positive correlation with the rates of both prevalence and YLDs in brain diseases including: Alzheimer's disease (AD) and brain cancer, but not in Parkinson's disease and depression. In particular, the spatial-temporal changes in tropospheric FA levels were consistent with the geographical distribution of FA exposure-induced AD and brain cancer in both sex old adults with age (60–89). Hence, mapping of tropospheric pollutants could be used for air quality monitoring and health risk assessment.

CP04.02 Dynamic Functional Connectivity Change-Point Detection with Random Matrix Theory Inference for Brain Network Data

Your name

Jaehee Kim

Abstract

Functional magnetic resonance imaging (fMRI) data is useful to study the dynamic nature of brain activity, including some temporal dependencies between the corresponding neural activity estimates. Modeling this dynamic functional connectivity (DFC) requires time-varying measures of the spatial region of interest (ROI) sets. Change-point detection in functional connectivity (FC) is particularly interesting for neurological inquiry and model flexibility. We propose detecting a change-point based on random matrix theory (RMT) with maximum eigenvalues. RMT approach is used for covariance matrices for DFC of all ROI's to decide the temporal change-point. Simulation results show that our proposed method can detect meaningful FC change-points.

We illustrate the effectiveness of our change-point detection method by applying it to epilepsy fMRI data investigating individual change-point. Our approach is exploratory with the inferential capabilities of the more rigid modeling approach. Our study shows the possibility of RMT based approach in studying the complex dynamic pattern of functional brain interactions.

CP03.03 Parametric Analysis of Bivariate Current Status data with Competing risks using Frailty models

Your name

Biswadeep Ghosh

Abstract

It can be noted that frailty variables have been used to model the association in bivariate failure time data with competing risks. For example, see Bandeen-Roche and Liang (2002), Gorfine and Hsu (2011). In this paper, we generalize the popular multiplicative frailty model using baseline cause-specific hazard and frailty variable possibly depending on the cause of failure, while focusing on bivariate current status data with competing risks. We have proposed a new parametric class of hazard function which has at least five different popular hazard functions as its special case. Four different Gamma frailty models have been considered where each model captures a different type of dependency. Relevant identifiability results corresponding to each of the above models have been proved. In order to investigate the finite sample behaviour of estimators, a detailed simulation study considering different sample sizes has been carried out. We have considered a real data analysis involving hearing disability using different parametric models. In this dataset, our newly proposed Gamma frailty models which are depending on competing risks are better fitted than the popular shared and correlated Gamma frailty models with respect to AIC.

CP03.04 On Limiting Behaviors of Stepwise Multiple Testing Procedures

Your name

Monitirtha Dey

Abstract

Stepwise multiple testing procedures have attracted several statisticians for decades and are also quite popular with statistics users because of their technical simplicity. The Bonferroni procedure has been one of the earliest and most prominent testing rules for controlling the familywise error rate (FWER). A recent article established that the FWER for the Bonferroni method asymptotically (i.e., when the number of hypotheses becomes arbitrarily large) approaches zero under any positively equicorrelated multivariate normal framework.

However, similar results for the limiting behaviors of FWER of general stepwise procedures are nonexistent. The present work addresses this gap in a unified manner by elucidating that, under the multivariate normal setups with certain correlation structures, the probability of rejecting one or more null hypotheses approaches zero asymptotically for any step-down procedure. Consequently, the FWER and Power of the step-down procedures also tend to be asymptotically zero. We also establish similar limiting zero results on FWER of other popular multiple testing rules, e.g., Hochberg's and Hommel's procedures. It turns out that, within our chosen asymptotic framework, the Benjamini-Hochberg method can hold the FWER at a strictly positive level asymptotically under the equicorrelated normality.

CP03.05 Statistical inferences for the F1-score of multi-label classification

Your name

Kanae Takahashi

Abstract

Data classification problems can be categorized into single-label classification and multi-label classification. In single-label classification, the data are mutually exclusive and are classified into exactly one of the classes. In multi-label classification, on the other hand, the data are not mutually exclusive and can be classified into several classes simultaneously. Several evaluation measures have been proposed for single-label and multi-label classifications. The F1-score, which is defined as the harmonic mean of precision (positive predictive value) and recall (sensitivity), is one of the evaluation measures used in both single-label and multi-label classifications. The F1-score is often used in the field of information retrieval and machine learning, and it is gaining popularity in the medical field. While There are interval estimation and hypothesis testing methods for the F1-score of single-label classification, point estimation can only be performed for F1-score of multi-label classification, and no interval estimation methods have yet been proposed. Therefore, this study proposes methods for estimating multi-label F1-score with confidence intervals based on the large-sample multivariate central limit theorem and the delta method. The performance of the proposed methods is investigated through simulations.

DL08: Distinguished Lecture Session

15:50 - 17:30 Friday, 5th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Kerrie Mengersen

Judith Rousseau: I Semi-parametric inference: A Bayesian curse?

Edwin Fong: Predictive Bayesian Inference with the Martingale Posterior

Susan Wei: Singular learning theory perspective on variational Bayesian neural networks

DL08.01 I Semi-parametric inference : A Bayesian curse?

Your name

Judith Rousseau

Abstract

In this talk I will discuss some issues around Bayesian approaches in semiparametric inference. I will first recall some positive and negative results on Bernstein von Mises theorems in non and semi-parametric models. I will then propose two possible tricks to derive \textit{posterior - type} distributions in semiparametric models which allow both for efficient procedures and

Bernstein von Mises theorems, as well as flexible priors on the nonparametric part. The first approach, based on the cut posterior will be illustrated in semi-parametric mixture and Hidden Markov models and second, a targeted posterior, will be applied in the well known causal inference problem of average treatment effect estimation.

DL08.02 Predictive Bayesian Inference with the Martingale Posterior

Your name

Edwin Fong

Abstract

While the prior distribution is the usual starting point for Bayesian uncertainty, recent work has reframed Bayesian inference as the predictive imputation of missing observations. In particular, the martingale posterior distribution arises when the Bayesian model is a chosen sequence of predictive distributions on future observables, which then induces a posterior distribution on the parameter of interest without the need for a likelihood and prior. This generalization greatly broadens the range of models one can use for Bayesian inference, and offers substantial advantages in computation and flexibility. In this talk, we introduce the framework and present some recent advances.

62 Singular learning theory perspective on variational Bayesian neural networks

Your name

Susan Wei

Abstract

In this work, we advocate for the importance of singular learning theory (SLT) as it pertains to the theory and practice of variational inference in Bayesian neural networks (BNNs).

To begin, we lay to rest some of the confusion surrounding discrepancies between downstream predictive performance measured via the test log predictive density and the variational objective. Next, we use the SLT-corrected asymptotic form for singular posterior distributions to inform the design of the variational family itself.

Specifically, we build upon the idealized variational family introduced in Bhattacharya et al. [2020] which is theoretically appealing but practically intractable. Our proposal takes shape as a normalizing flow where the base distribution is a carefully-initialized generalized gamma. We conduct experiments comparing this to the canonical Gaussian base distribution and show improvements in terms of variational free energy and variational generalization error.

IP39: "Statistical inference for branching processes"

15:50 - 17:30 Friday, 5th January, 2024
B305 - Room 305, Babel Building
Chairs Sophie Hautphenne
Organiser: Peter Braunsteins

Carmen Minuesa: ABC methodology for inference on the maximal offspring and parameters in controlled branching processes

Peter Braunsteins: Consistent estimation for population-size-dependent branching processes

James Kerlidis: Linking Population-Size-Dependent and Controlled Branching Processes

Sophie Hautphenne: Consistent estimation in subcritical birth-and-death processes

IP39.01 ABC methodology for inference on the maximal offspring and parameters in controlled branching processes

Your name

Carmen Minuesa Abril

Abstract

In a Bayesian framework, our aim is to estimate the posterior distribution of the parameters of controlled branching processes without computing the likelihood function and when the maximum number of children that an individual can give birth to is unknown.

Our approach involves adapting and implementing approximate Bayesian computation (ABC) methods in the context of branching processes. We present a rejection ABC algorithm for model selection that enables us to estimate the maximum number of offspring per individual. To that end, we compare the raw simulated sample with the observed sample. The second step consists in approximating the posterior distributions of the parameters using a tolerance-rejection algorithm and a post-sampling correction method, along with a suitable summary statistic.

We illustrate the accuracy of the proposed methods by means of a simulated example developed with the statistical software R. We also apply our results to a real dataset of harbour seals.

This is a joint work with Miguel González and Inés del Puerto.

IP39.02 Consistent estimation for population-size-dependent branching processes

Your name

Peter Braunsteins

Abstract

In this talk we consider population-size-dependent branching processes, which are stochastic models designed to capture the dynamics of populations living in restricted habitats. We derive several estimators for parameters in these processes and demonstrate that some of these estimators do not satisfy the classical consistency property called C-consistency. This leads us to define a new type of consistency called Q-consistency. We then investigate the circumstances when C-consistent estimators are preferable to Q-consistent estimators. The results are illustrated on the Chatham Island black robin population.

IP39.03 Linking Population-Size-Dependent and Controlled Branching Processes

Your name

James Kerlidis

Abstract

Biological populations exposed to phenomena such as resource scarcity and immigration exhibit dynamics more complex than can be modelled with the classical Galton-Watson branching process (GWBP). Population-size dependent branching processes (PSDBPs) and controlled branching processes (CBPs) both distinctly extend the GWBP to allow for such complex dynamics, and both are popular models for biological populations that exhibit logistic growth.

We may wish to determine if a given population is more appropriately modelled with a PSDBP or a CBP. We consider the dual of this question, and investigate populations where a modeller would be indifferent to the choice. In particular, we state conditions for the existence of equivalent PSDBPs and CBPs, and derive an upper bound on the total variation distance between non-equivalent PSDBPs and CBPs with matching mean and variance and equal initial population size, and show that under certain conditions this bound tends to zero as the initial population size becomes large.

IP39.04 Consistent estimation in subcritical birth-and-death processes

Your name

Sophie Hautphenne

Abstract

Birth-and-death processes are the simplest continuous-time branching processes and are widely used in applications. Subcritical birth-and-death processes model declining populations that would naturally die out without conservation efforts. When we apply these processes to model endangered populations, the data should be interpreted as generated by the process conditioned on survival, and consistent estimators are desirable.

In this talk, we propose the first consistent estimators for the birth and death rates of subcritical birth-and-death processes. Our estimators are based on continuous observation of a single non-extinct trajectory of the process. The idea behind their construction stems from a path-wise decomposition of subcritical branching processes conditioned to survive in the distant future. This is joint work with Emma Horton.

IP26: "Recent developments on combinatorial stochastic processes"

15:50 - 17:30 Friday, 5th January, 2024
OA239 - Room 239, Old Arts Building
Chairs Takeru Matsuda
Organiser: Takeru Matsuda

Nobuaki Hoshino: A Bell polynomial process

Shuhei Mano: A measure-on-graph-valued diffusion: a particle system with collisions and its applications

IP26.01 A Bell polynomial process

Your name

Nobuaki Hoshino

Abstract

Any margin of the multinomial distribution is multinomially distributed. Retaining this closure, we construct a family of generalized multinomial distributions. This family is characterized within multiplicative probability measures, using the Bell polynomial. The family actually includes the mixed multinomial distribution with the normalized infinitely divisible distribution. Concerning this, the family can be obtained by conditioning independent compound-Poisson-distributed variables on the sum, and it defines a finite point process, being regarded as the finite dimensional distribution. Due to the closure under taking a margin, the family is very treatable in practices. Some finite dimensional properties such as marginal moments will be presented, as well as a conjugacy of this family for Bayesian analyses.

IP26.02 A measure-on-graph-valued diffusion: a particle system with collisions and its applications

Your name

Shuhei Mano

Abstract

A diffusion process taking value in probability measures on vertices of graphs is studied. The masses on each vertex satisfy a stochastic differential equation on the simplex. A dual Markov chain on integer partitions to the Markov semigroup associated with the diffusion is used to show that the support of an extremal stationary state of the adjoint semigroup is an independent set of the graph. We also investigate the diffusion with a linear drift, which gives a killing of the dual Markov chain on a finite integer lattice. The Markov chain is used to study the unique stationary state of the diffusion, which generalizes the Dirichlet distribution. Two applications of the diffusions are discussed: analysis of an algorithm to find an independent set of a graph and a Bayesian graph selection based on a computation of the probability of a sample by using coupling from the past.

IP43: "Emerging opportunities in omics data and longitudinal/functional data analysis"

15:50 - 17:30 Friday, 5th January, 2024
OA107 - Room 107, Old Arts Building
Chairs Ke Deng
Organiser: Di Wu

Kun Huang; Identify Alzheimer's disease subtypes and markers from multi-omic data of human brain and blood using subspace merging approach

Fei Zou: Advanced Learning Algorithms for Genetics and Genomics Data

Ying Zhang: clusterMLD: An Efficient Clustering Method for Multivariate Longitudinal Data

Ke Deng: Model-Based Spatial Reconstruction of Large-Scale Biomolecules via Bayesian Inference of a Hierarchical Spatial Model

IP43.01 Identify Alzheimer's disease subtypes and markers from multi-omic data of human brain and blood using subspace merging approach

Your name

Kun Huang

Abstract

Alzheimer's disease (AD) is a highly heterogeneous disease with different disease trajectory, brain pathological changes, and risk factors. Identifying subtypes of AD can have a big impact on AD diagnosis, treatment, and future disease management. In this project we applied an unsupervised subspace merging algorithm on AD patient brain samples with matched transcriptomics, DNA methylation, and proteomics data collected from the Religious Orders Study and Memory and Aging Project (ROSMAP). This algorithm merges patient graphs generated from these three data types on a Grassmann manifold followed by unsupervised spectral clustering. We identified two patient clusters in which the cognitive test scores or the degree of pathology in AD are significantly different. Differentially gene expression and downstream pathway analysis identified gene signatures between the patient clusters as well as to the transcriptomic data of the peripheral blood monocyte (PBMC) from the matching patients to check for possible blood markers. Ingenuity Pathway Analysis was applied to examine any connections between the DEGs in the brain samples with the ones identified from the PBMC samples. Among the DEGs, an apolipoprotein E receptor gene identified from the blood sample is associated with 12 DEGs from the brain samples, suggesting a potential novel blood biomarker for identifying AD subtypes.

IP43.02 Advanced Learning Algorithms for Genetics and Genomics Data

Your name

Fei Zou

Abstract

Deep Neural Network has become one of the most popular machine learning algorithms in biomedical research due to its high flexibility in approximating complex functions. In this talk, we will present a deep neural network based algorithm for personalized drug response prediction and a computational pipeline for identifying sequences of antigens that stimulate immune response (i.e., peptide presentation). In addition, a permutation-based feature importance scoring technique that facilitates the identification of causal features even in high-collinearity scenarios, enabling accurate biomarker identifications and model interpretations for these seemingly "black-box" models will be discussed.

IP43.03 clusterMLD: An Efficient Clustering Method for Multivariate Longitudinal Data

Your name

Ying Zhang

Abstract

Longitudinal data clustering is a challenging task, especially with sparse and irregular observations. It lacks reliable methods in the literature that deal with clustering complicated longitudinal data, particularly with multiple longitudinal outcomes. In this manuscript, a new agglomerative hierarchical clustering method is developed in conjunction with B-spline curve fitting and construction of unique dissimilarity measure for differentiating longitudinal observations. In an extensive simulation study, the proposed method demonstrates its superior performance in clustering accuracy and numerical efficiency compared to the existing methods. Moreover, the method can be easily extended to multiple-outcome longitudinal data without too much cost in computation and shows its robust results against the complexity of underlying mixture of longitudinal data. Finally, the method is applied to a date set from the SPRINT Study for validating the intervention efficacy in a Systolic Blood Pressure Intervention Trial and to a 12-year multi-site observational study (PREDICT-HD) for identifying the disease progression patterns of Huntington’s disease (HD).

IP43.04 Model-Based Spatial Reconstruction of Large-Scale Biomolecules via Bayesian Inference of a Hierarchical Spatial Model

Your name

Ke Deng

Abstract

Revealing the spatial organization of biomolecules and characterizing their spatial distribution in cells and tissues have long been recognized as importance problems in biomedical research. With rapid advances of DNA sequencing technologies in recent years, creative sequencing-based experimental assays, e.g., Hi-C and DNA microscopy, have been invented to reveal the spatial properties of large-scale biomolecules in a high-throughput and high-resolution manner. A typical experiment based on these technologies produces a count matrix to record the contact frequencies among molecules of interest, which are closely associated to their spatial distances, allowing us to reconstruct the spatial organization of large-scale biomolecules via data analysis. There is a great appeal to develop statistically rigorous and computationally scalable methods for this important problem. In this study, we fill in this gap with a novel method named HiSpa. Equipped with a hierarchical spatial model, HiSpa utilizes the idea of multi-scale modelling to reduce the computation complexity from O(n2) to O(n3/2) with little loss on the quality of the reconstructed spatial structure. Advanced Monte Carlo strategies are developed for efficient Bayesian inference of HiSpa. Superiority of HiSpa over existing methods is demonstrated by simulation studies and real data applications.

IP46: "New Techniques for Analyzing Big Data"

15:50 - 17:30 Friday, 5th January, 2024
AW153 - Room 153, Arts West North Building
Chairs Juan Hu
Organiser: Yichao Wu

Juan Hu: A Unified Approach to Variable Selection for Partially Linear Models

Juhyun Park: Geometric functional data analysis for multivariate curves

Peter Radchenko: A Discrete Optimization Approach to Learning with Categorical Variables

Cun-Hui Zhang: Adaptive Inference in Sequential Experiments

IP46.01 A Unified Approach to Variable Selection for Partially Linear Models

Your name

Juan Hu

Abstract

We focus on the general partially linear model without any structure assumption on the nonparametric component. For such a model with both linear and nonlinear predictors being multivariate, we propose a new variable selection method. Our unique method is a unified approach in that it can select both linear and nonlinear predictors simultaneously by solving a single optimization problem. We prove that the proposed method achieves consistency. Both simulation examples and a real data example demonstrate the new method's competitive finite-sample performance.

IP46.02 Geometric functional data analysis for multivariate curves

Your name

Juhyun Park

Abstract

Increasing number of sensors collects a huge volume of complex data. Of particular interest is data in the form of multivariate curves, which may represent complex signals from physical or biomechanical experiments, movements or object tracking. These types of data are challenging to analyze because they are meaningful but are difficult to characterize without domain knowledge.

Functional data analysis is well developed for the analysis of curves, focusing on identifying commonalities and quantifying variations, typically based on the analytical representation of functions. On the other hand, the form or shape is one of the essential features of objects that help us understand and characterize them. Studying geometric features of the objects, independently of the underlying coordinate system, has been the subject of shape analysis. A well know example to incorporate geometric features into functional data analysis is the study of phase and amplitude variation of the curves. However, its extension to multivariate curves is not straightforward.

In this work, we study the characterization of the geometric features of multivariate curves. We utilize a Riemannian framework of differential geometry to define a metric between the shapes to compare. Under this framework, the choice of the metric needs careful consideration, to be adaptive to the types of data. We propose new geometric representations of the curves, which will be the basis of statistical shape analysis for multivariate curves. Examples are used to illustrate how different choices of the metric are linked to the representation of the shapes of the curves.

IP46.03 A Discrete Optimization Approach to Learning with Categorical Variables

Your name

Peter Radchenko

Abstract

We will consider a high-dimensional linear regression problem where most of features are categorical, each having a large number of levels. We will discuss an estimator that fuses some of the levels of each variable together by making their corresponding coefficients exactly equal. This is achieved by directly penalizing the total number of levels in the regression coefficients. The proposed estimator also controls the overall sparsity of the solution via an additional L0 penalty. We will show that the proposed estimator can be written as a solution to a mixed-integer programming problem, where a quadratic objective is minimized over mixed-integer linear constraints. Thus, the estimator is amenable to modern computational tools from mathematical optimization. We will discuss approximate algorithms for our method based on block coordinate descent that obtain high-quality feasible solutions in a computationally efficient fashion. We will explore the statistical properties of the proposed estimator and demonstrate that it may have an edge over existing approaches.

IP46.04 Adaptive Inference in Sequential Experiments

Your name

Cun-Hui Zhang

Abstract

Sequential data collection has emerged as a widely adopted technique for enhancing the efficiency of data gathering processes. Despite its advantages, such data collection mechanism often introduces complexities to the statistical inference procedure. For instance, the ordinary least squares estimator in an adaptive linear regression model can exhibit non-normal asymptotic behavior, posing challenges for accurate inference and interpretation. We propose a general method for constructing debiased estimator which remedies this issue. The idea is to make use of adaptive linear estimating equations. We establish theoretical guarantees of asymptotic normality, supplemented by discussions on achieving near-optimal asymptotic variance. This talk is based on joint work with Mufang Ying and Koulik Khamaru.

IP52: "Advances in inference for multivariate and high dimensional data"

15:50 - 17:30 Friday, 5th January, 2024
B106 - Room 106, Babel Building
Chairs Aurore Delaigle
Organiser: Aurore Delaigle

Karim Seghouane: Joint Approximate Partial Diagonalization of Large Matrices

Debashis Paul: Spectral behavior of sequential sample covariances under a spiked model

Linh Nghiem: Random efffect sufficient dimension reduction for clustered data

Gery Geenens: Universal copulas

IP52.01 Joint Approximate Partial Diagonalization of Large Matrices

Your name

Karim Seghouane

Abstract

Given a set of p symmetric (real) matrices, the Orthogonal Joint Diagonalization (OJD) problem consists of finding an orthonormal basis in which the representation of each of these p matrices is as close as possible to a diagonal matrix. We argue that when the matrices are of large dimension, then the natural generalization of this problem is to seek an orthonormal basis of a certain subspace that is a near eigenspace for all the matrices in the set. We refer to this as the problem of “partial joint diagonalization of matrices.” The approach proposed first finds this approximate common near eigenspace and then proceeds to a joint diagonalization of the restrictions of the input matrices in this subspace. A few solution methods for this problem are proposed and illustrations of its potential applications are provided.

IP52.02 Spectral behavior of sequential sample covariances under a spiked model

Your name

Debashis Paul

Abstract

Suppose we have i.i.d. observations, indexed by time, from a multivariate distribution. The corresponding *sequential sample covariance matrices*, defined as the sample covariance matrix computed from observations up to time *t*, form a matrix-valued stochastic process indexed by time. We study the spectral statistics of such matrices as stochastic processes, when the dimension of the data vectors is comparable to the sample size. First results of this kind have recently been established by Dörnemann and Dette (2023), who derived the asymptotic distribution of linear spectral statistics associated with the sequential covariance matrices. We study a related problem, that of characterizing the behavior of normalized spiked eigenvalues of the sequential covariances when the population covariance matrix is a finite rank perturbation of the identity matrix, and the spiked eigenvalues are sufficiently big. We establish the existence of a limiting process associated with the normalized eigenvalues. We also consider an application of the results to constructing rotation-invariant tests for detecting change points in the structure of the population covariance matrix when such changes only affect the sizes of the spiked eigenvalues.

(This is a joint work with Nina Dörnemann).

IP52.03 Random efffect sufficient dimension reduction for clustered data

Your name

Linh Nghiem

Abstract

Sufficient dimension reduction (SDR) is a popular class of regression methods aiming to find a small number of linear combinations of covariates that capture all the information of the responses (a central subspace). The majority of current methods for SDR focus on the setting of independent observations, while the few SDR techniques developed for clustered data assume the linear transformation is identical across clusters. In this article, we introduce the concept of random effect central subspaces, where cluster-specific central subspaces are assumed to be random following a distribution on the Grassmann manifold.

This random effects distribution is the image of an exponential mapping from a Gaussian distribution on the tangent space of an overall fixed effects central subspace, and characterized by a covariance matrix capturing the heterogeneity between clusters. We incorporate the random effect central subspaces into a principal fitted components model, and propose a two-stage algorithm for estimation and prediction of the cluster-specific central subspaces. We demonstrate the consistency of the proposed estimators when the number of clusters grows while the cluster sizes remain bounded. Simulation studies demonstrate the superior performance of our proposed approach compared to both global and cluster-specific SDR methods. We apply the proposed method to study the relationship between the life expectancy of women with socioeconomic variables across countries. Results show log income per capita, infant mortality, and inequality primarily drive the overall fixed effects central subspace, although there is considerable variability between countries in how their cluster-specific central subspaces are driven by these predictors.

IP52.04 Universal copulas

Your name

Gery Geenens

Abstract

The early 21st century has seen the advent of copulas as primary statistical tools when it comes to model dependence between numerical random variables. A copula is classically understood as a cumulative distribution function on the unit hypercube with standard uniform margins – we refer to such distributions as “Sklar’s copulas”, owing to their central role in the decomposition of multivariate distributions established by the celebrated Sklar's theorem. The argument habitually put forward for outlining the appeal of copula models is that they allow pulling apart the dependence structure of a bivariate vector, characterised by its copula, from the individual behaviour of its marginal components. Though, this interpretation can only be justified in the continuous framework, as copulas lose their "margin-free" nature outside of it, making Sklar’s copula models unfit for modelling dependence between non-continuous variables.

In this work, we argue that the very notion of copula should not be imprisoned into Sklar's theorem, and we propose an alternative definition of copulas which follows from approaching their role and meaning more broadly. This definition coincides with Sklar's copulas in the continuous framework, but leads to different concepts in other settings.

We call this construction “universal copulas” and show that these maintain in all situations the pleasant properties (in particular: "margin-freeness") which make Sklar’s copulas sound and effective in the continuous case. We illustrate our findings with some examples of “universal copula modelling” between two discrete variables, and between one continuous variable and one Bernoulli variable.

IP68: "Recent Advances in Statistical and Machine Learning"

15:50 - 17:30 Friday, 5th January, 2024
EEBrown - Brown Room, Electrical Engineering Building
Chairs Jianqing Fan
Organiser: Jianqing Fan

Johannes Schmidt-Heiber: Statistical learning in biological neural networks

Michael Kohler: Learning of deep network classifiers via gradient descent and over-parametrization

Haofeng Wang: Empirical likelihood ratio tests for non-nested model selection based on predictive losses

IP68.01 Statistical learning in biological neural networks

Your name

Johannes Schmidt-Hieber

Abstract

Compared to artificial neural networks (ANNs), the brain learns faster, generalizes better to new situations and consumes much less energy. ANNs are motivated by the functioning of the brain, but differ in several crucial aspects. For instance, ANNs are deterministic while biological neural networks (BNNs) are stochastic. Moreover, it is biologically implausible that the learning of the brain is based on gradient descent. In this talk we look at biological neural networks as a statistical method for supervised learning. We relate the local updating rule of the connection parameters in BNNs to a zero-order optimization method and derive some first statistical risk bounds.

IP68.02 Learning of deep network classifiers via gradient descent and over-parametrization

Your name

Michael Kohler

Abstract

Pattern recognition from independent and identically distributed random variables is considered. A general result is presented which shows how a deep network classifier can be fitted to data via over-parametrization and gradient-descent. The result is used to analyze the rate of convergence of the difference of the misclassification probability of a classifier based on a (deep) Transformer encoder and the optimal misclassification probability in case that the aposteriori probability satisfies a smooth hierarchical composition model.

IP68.03 Empirical likelihood ratio tests for non-nested model selection based on predictive losses

Your name

Haofeng Wang

Abstract

We propose an empirical likelihood ratio (ELR) test for comparing any two supervised learning models, which may be nested, non-nested, overlapping, misspecifified, or correctly specified. The test compares the prediction losses of models based on the cross-validation. We determine the asymptotic null and alternative distributions of the ELR test for comparing two nonparametric learning models under a general framework of convex loss functions. However, the prediction losses from the cross-validation involve repeatedly fitting the models with one observation left out, which leads to a heavy computational burden. We introduce an easy-to-implement ELR test which requires fifitting the models only once and shares the same asymptotics as the original one. The proposed tests are applied to compare additive models with varying-coeffificient models. Furthermore, a scalable distributed ELR test is proposed for testing the importance of a group of variables in possibly misspecifified additive models with massive data. Simulations show that the proposed tests work well and have favorable finite-sample performance compared to some existing approaches. The methodology is validated on an empirical application.

CP12: Contributed Paper Session

15:50 - 17:30 Friday, 5th January, 2024
BG03 - Room G03, Babel Building
Chairs Jiangrong Ouyang

Anik Roy: A Control Chart For Online Shape Monitoring In Image Data

Chieh-An Chou: Finite Difference Spatial Spline for Big Data

Yulin Zhang: Semiparametric Estimation of Continuous Treatment Effect on the Treated by Balancing Index Moments

Masayuki Henmi: Infinite-dimensional information geometry for semiparametric statistics

Jiangrong Ouyang: Bayesian analysis of longitudinal data via empirical likelihood

CP12.01 A Control Chart For Online Shape Monitoring In Image Data

Your name

Anik Roy

Abstract

Sequential monitoring of the shape of the image object is a relatively new research area in statistics and machine learning. It has a wide range of applications in different disciplines of science, including satellite imaging, medical research, industrial applications, and so forth. For instance, the gradual shrinking of the Salton Sea comprises the loss of water resources, damaging the local environment and ecosystems. Note that shape monitoring in the image data is a challenging problem because (i) image surface contains discontinuities in the boundary of the image object; therefore, the traditional smoothing technique is inefficient, (ii) shape of the image object is often irregular in nature, and (iii) sequence of images is not always geometrically aligned; therefore, the problem of rigid body image registration is also associated with it. In this article, we propose a shape monitoring algorithm that is rotation and translation invariant, therefore capable of detecting changes in the shape of the image object efficiently while ignoring the change due to rigid body image transformation.

By this method, for comparing two shapes (two images), we construct a test statistic based on the distribution of the distance from the centroid of the image object to the boundary. For online monitoring, we propose a nonparametric CUSUM control chart based on the statistic mentioned above. The proposed method is simple to interpret and also capable of handling all the issues mentioned above. Theoretical justification and numerical performances show that the proposed algorithm works well in many different situations.

CP12.02 Finite Difference Spatial Spline for Big Data

Your name

Chieh-An Chou

Abstract

The advent of big data has created a demand for models capable of efficiently handling large datasets. Spatial data analysis, in particular, presents additional challenges due to its incorporation of multivariate elements. These challenges are further compounded by the presence of terrain constraints, such as islands, which require the problem to be defined within a finite and intricate domain, resulting in the issue of “irregular domain”.

Traditional spatial methods, such as thin plate splines, suffer from high computational complexity and lack of consideration for domain-specific errors. Although methods like soap film smoothing have made improvements in addressing irregular domains, there is still a lack of models that can be applied to big data.

This thesis introduces the Finite Difference Spatial Spline model (FDSS) to overcome these challenges. FDSS directly estimates function values on a discretized domain and optimizes them using a finite difference approach, thereby eliminating the need for complex model matrices and enabling faster fitting procedures for large datasets. It also demonstrates versatility by effectively handling irregular domains.

In the simulation experiments, FDSS demonstrated impressive performance in both regular and irregular domains. It was able to handle datasets with millions of data points and complete computations within a matter of seconds. Despite its speed, FDSS maintained accuracy comparable to other models.

CP12.03 Semiparametric Estimation of Continuous Treatment Effect on the Treated by Balancing Index Moments

Your name

Yulin Zhang

Abstract

Continuous treatments arise commonly in practice and have gained increasing attention in recent years, but most of the existing literature focuses on the population average dose-response function (ADRF). This paper proposes a univariate parametric model for the average dose-response function on the treated group (ADRFT), which defines the treatment effects of a continuous treatment on a subpopulation. To get rid of the ''curse of dimensionality", we consider a single-index model for the generalized propensity score (GPS) and derive the semiparametric efficiency bound for the ADRFT model under the unconfoundedness. We show that the ADRFT can be identified through a weighted least square regression of the response on the treatment variable, where the weights is a function of the inverse GPS. We propose a class of estimators for the weighting function by balancing an increasing number of single-index moments. The root-n consistency and asymptotic normality of the proposed estimators for the ADRFT model are established, which is shown to attain the semiparametric efficiency bound if the outcome regression (OR) function is also single-indexed. Monte-Carlo simulations and a real data application demonstrate the practical value of the proposed method.

CP12.04 Infinite-dimensional information geometry for semiparametric statistics

Your name

Masayuki Henmi

Abstract

In information geometry, a parametric statistical model (a parametric family of probability density functions) is treated as a differentiable manifold, where the Riemannian metric called Fisher metric and the pair of two torsion-free dual affine connections called the exponential and mixture connections play essential roles for geometrical interpretation of statistical inference.

For example, the maximum likelihood estimation in an exponential family is viewed as the orthogonal projection of the geodesic defined by the mixture connection and its statistical properties can be understood in terms of geometry. However, semi-parametric or non-parametric statistical methods cannot be understood geometrically in this framework because the statistical models behind these methods are essentially infinite-dimensional. Since Pistone and Sempi (1995) introduced an infinite-dimensional differentiable manifold of probability distributions using the notion of Orlicz spaces, the geometrical theory of infinite-dimensional statistical models has been highly developed from a mathematical point of view. However, it has not sufficiently linked to geometrical understanding of semi-parametric or non-parametric statistical methods yet. In this presentation, we summarize the previous (mathematical) studies and clarify the problems for the development in statistics. In particular, we focus on the semiparametric estimation theory and consider the possibility to develop it as a geometrical theory in which the semiparametric model is viewed as an infinite-dimensional differential manifold.

CP12.05 Bayesian analysis of longitudinal data via empirical likelihood

Your name

Jiangrong Ouyang

Abstract

Longitudinal data consists of repeated observations that are typically correlated, which makes the likelihood-based inference challenging. This limits the use of Bayesian methods for longitudinal data in many general situations. To address this issue, empirical likelihood is used to develop a fully Bayesian method for analyzing longitudinal data based on a set of moment equations parallel to the form of generalized estimating equations. It is demonstrated in the context of two popular priors for Bayesian inference and regularization, the Laplace prior and the horseshoe prior. The proposed Bayesian shrinkage method performs well in both estimation accuracy and variable selection, while also providing a full quantification of uncertainty. The method is illustrated using a yeast cell-cycle microarray time course gene expression data set.

DL07: Distinguished Lecture Session

08:30 - 10:10 Saturday, 6th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Jialiang Li

Liza Levina: Latent space models for multiplex networks

Tianxi Li: Network comparison by multivariate moment inference

Jesus Arroyo: Learning Joint and Individual Structure in Network Data with Covariates

DL07.01 Latent space models for multiplex networks

Your name

Liza Levina

Abstract

Statistical tools for analysis of a single network are now widely available, but many practical settings involve multiple networks. These can arise as a sample of networks (for example, brain connectivity networks for a sample of patients), a single network with multiple types of edges (for example, trade between countries in many different commodities), or a single network evolving over time. The term multiplex networks refers to multiple and generally heterogeneous networks observed on the same shared node set; the two examples above are both multiplex networks. We propose a new latent space model for multiplex networks which answers a key question: what part of the underlying structure is shared between all the networks, and what is unique to each one? Our model learns this from data and pools information adaptively. We establish identifiability, develop a fitting procedure using convex optimization in combination with a nuclear norm penalty, and prove a guarantee of recovery for the latent positions as long as there is sufficient separation between the shared and the individual latent subspaces. We compare the model to competing methods in the literature on simulated networks and on a multiplex network describing the worldwide trade of agricultural products. This is joint work with Peter MacDonald and Ji Zhu.

DL07.02 Network comparison by multivariate moment inference

Your name

Tianxi Li

Abstract

Comparing networks of varying sizes is a pivotal statistical challenge, crucial for interpreting complex network data across fields such as biology, medicine, and social sciences. A robust comparison technique must be both statistically rigorous and produce results that are scientifically insightful. This presentation delves into comparing networks through multivariate inference of network moments. We will illustrate that single-moment inference is insufficient for deriving meaningful conclusions; it's imperative to incorporate both joint and conditional inference. We will adapt the subsampling technique to multivariate contexts, showcasing its potency in yielding more discernible test outcomes. The method will be illustrated by comparison studies of several real-world networks.

DL07.03 Learning Joint and Individual Structure in Network Data with Covariates

Your name

Jesus Arroyo

Abstract

Datasets consisting of a network and covariates associated with its vertices have become ubiquitous. One problem with this type of data is to identify information unique to the network, information unique to the vertex covariates, and information that is shared between the network and the vertex covariates. Existing methods for network data often focus on capturing structure that is shared between a network and the vertex covariates but are not able to differentiate structure that is unique to each. This work formulates a solution via a low-rank model and a two-step estimation procedure composed of an efficient spectral method to obtain an initial estimate for the joint structure, followed by an optimization method that minimizes a nonconvex loss function associated with the model. We study the consistency of the initial estimate and evaluate the performance on simulated and real data.

IP24: "Statistical inference for non-standard data and complex models"

08:30 - 10:10 Saturday, 6th January, 2024
AW153 - Room 153, Arts West North Building
Chairs Kyusang Yu
Organiser: Young Kyung Lee

Ming-Yen Cheng: Statistical analysis of dependent, high dimensional and massive data

Jeong Min Jeon: Density estimation on Lie groups in the presence of measurement error without auxiliary data

Kyunghee Han: Testing linear operator constraints in functional response regression with incomplete response functions

Seong Jun Yang: Cure models with time-varying coefficients in hazards

IP24.01 Inference for changing periodicity, trend and covariate effects in nonstationary time series

Your name

Ming-Yen Cheng

Abstract

Motivated by two examples concerning global warming and monthly total import and export by China, we study time series that contain a nonparametric periodic component with an unknown period, a nonparametric trending behavior and also additive covariate effects. Further, as the amplitude function may change at some known or unknown change-point(s), we extend our model to take this dynamical periodicity into account and introduce two change-point estimators. To the best of knowledge, this is the first work to study such complex periodic structure. A two-step estimation procedure is proposed to estimate accurately the periodicity, trend and covariate effects. First, we estimate the period with the trend and covariate effects being approximated by B-splines rather than being ignored. To achieve robustness we employ a penalized M-estimation method which utilizes post model selection inference ideas. Next, given the period estimate, we estimate the amplitude, trend and covariate effects. Asymptotic properties of our estimators are derived, including consistency of the period estimator and asymptotic normality and oracle property of the estimated periodic sequence, trend and covariate effects. Simulation studies confirm superiority of our method and illustrate good performance of our change-point estimators. Applications to the two motivating examples demonstrate utilities of our methods.

IP24.02 Density estimation on Lie groups in the presence of measurement error without auxiliary data

Your name

Jeong Min Jeon

Abstract

In this talk, we introduce density estimation on a general Lie group when data contain measurement errors and the distribution of measurement error is unknown. We estimate the target density without additional observations such as an observable random sample from the measurement error distribution or repeated measurements. To achieve this, we take a semiparametric approach, which assumes that the measurement error distribution belongs to a parametric family. We also investigate a fully parametric approach for the case where the target density is also parametric. We establish an identifiability result for a measurement error model on the general Lie group and derive various asymptotic properties of our density estimators. Simulation studies are performed to demonstrate the superior performance of our estimators.

IP24.03 Testing linear operator constraints in functional response regression with incomplete response functions

Your name

Kyunghee Han

Abstract

Hypothesis testing procedures are developed to assess linear operator constraints in function-on-scalar regression when incomplete functional responses are observed. The approach enables statistical inferences about the shape and other aspects of the functional regression coefficients within a unified framework encompassing three incomplete sampling scenarios: (i) partially observed response functions as curve segments over random sub-intervals of the domain, (ii) discretely observed functional responses with additive measurement errors, and (iii) the composition of former two scenarios, where partially observed response segments are observed discretely with measurement error. The latter scenario has been little explored to date, although such structured data is increasingly common in applications.

For statistical inference, deviations from the constraint space are measured via integrated L^2-distance between estimates from the constrained and unconstrained model spaces. Large sample properties of the proposed test procedure are established, including the consistency, asymptotic distribution, and local power of the test statistic.

IP24.04 Cure models with time-varying coefficients in hazards

Your name

Seong J. Yang (Gwangsu Kim)

Abstract

Cure models take into account the situation in which some of the subjects may not experience the event of interest. In the analysis of time-to-event data, these models can be valuable not only when it is assumed that there are individuals who are actually cured from a certain disease but also when there are many long-term survivors. This study introduces a promotion time cure model, closely linked to the Cox model, with incorporating time-varying coefficients when modeling the effect of covariates. This consideration adds flexibility to the model by addressing non-proportionality and allows for the assessment of time-varying effects in the hazard function. Estimators for the coefficients under the introduced model are proposed, and their asymptotic normalities are discussed. The reliability of the proposed estimation procedure is also examined through numerical experiments.

IP13: "Statistical Analysis of Network Data"

08:30 - 10:10 Saturday, 6th January, 2024
B305 - Room 305, Babel Building
Chairs Binyan Jiang

Ruijian Han: A unified analysis of likelihood-based estimators in the Plackett–Luce model

Tao Zou: A Mutual Influence Model for Two-Mode Network Data

Fangyi Wang: Distribution-Free Matrix Prediction Under Arbitrary Missing Pattern

Binyan Jiang: A two-way heterogeneity model for dynamic networks

IP13.01 A unified analysis of likelihood-based estimators in the Plackett–Luce model

Your name

Ruijian Han

Abstract

The Plackett–Luce model is a popular approach for ranking data analysis, where a utility vector is employed to determine the probability of each outcome based on Luce's choice axiom. In this work, we investigate the asymptotic theory of utility vector estimation by maximizing different types of likelihood, such as the full-, marginal-, and quasi-likelihood. We provide a rank-matching interpretation for the estimating equations of these estimators and analyze their asymptotic behavior as the number of items being compared tends to infinity. In particular, we establish the uniform consistency of these estimators under conditions characterized by the topology of the underlying comparison graph sequence and demonstrate that the proposed conditions are sharp for common sampling scenarios such as the nonuniform random hypergraph model and the hypergraph stochastic block model; we also obtain the asymptotic normality of these estimators and discuss the trade-off between statistical efficiency and computational complexity for practical uncertainty quantification. Both results allow for nonuniform and inhomogeneous comparison graphs with varying edge sizes and different asymptotic orders of edge probabilities. We verify our theoretical findings by conducting detailed numerical experiments.

IP13.02 A Mutual Influence Model for Two-Mode Network Data

Your name

Tao Zou

Abstract

A two-mode network is a distinctive network structure where nodes are divided into two specific categories, and edges exclusively connect nodes of differing types. These network connections often lead to interdependencies between nodes of one type and nodes of the other type, and the nature of these relationships can vary across nodes. To examine and model this heterogeneity, we introduce a mutual influence model tailored for two-mode network data. In this model, we account for node-specific heterogeneity by incorporating two sets of influence parameters. The first set of parameters gauges the extent of influence each node exerts on others, while the second set of parameters quantifies how receptive each node is to being influenced by others. To estimate this model, we introduce the quasi-maximum likelihood estimator and establish its asymptotic properties. To evaluate the performance of our proposed model and estimation in finite samples, we conduct simulation studies and provide an empirical example.

IP13.03 Distribution-Free Matrix Prediction Under Arbitrary Missing Pattern

Your name

Fangyi Wang

Abstract

This paper studies the open problem of conformalized entry prediction in a row/column-exchangeable matrix. The matrix setting presents novel and unique challenges, but there exists little work on this interesting topic. We meticulously define the problem, differentiate it from closely related problems, and rigorously delineate the boundary between achievable and impossible goals. We then propose two practical algorithms. The first method provides a fast emulation of the full conformal prediction, while the second method leverages the technique of algorithmic stability for acceleration. Both methods are computationally efficient and can effectively safeguard coverage validity in presence of arbitrary missing pattern. Further, we quantify the impact of missingness on prediction accuracy and establish fundamental limit results. Empirical evidence from synthetic and real-world data sets corroborates the superior performance of our proposed methods.

IP13.04 A two-way heterogeneity model for dynamic networks

Your name

Binyan Jiang

Abstract

Analysis of networks that evolve dynamically requires the joint modelling of individual snapshots and time dynamics. This paper proposes a new flexible two-way heterogeneity model towards this goal. The new model equips each node of the network with two heterogeneity parameters, one to characterize the propensity to form ties with other nodes statically and the other to differentiate the tendency to retain existing ties over time. With n observed networks each having p nodes, we develop a new asymptotic theory for the maximum likelihood estimation of 2p parameters when np--> infinity. We overcome the global non-convexity of the negative log-likelihood function by the virtue of its local convexity, and propose a novel method of moment estimator as the initial value for a simple algorithm that leads to the consistent local maximum likelihood estimator (MLE). To establish the upper bounds for the estimation error of the MLE, we derive a new uniform deviation bound, which is of independent interest. The theory of the model and its usefulness are further supported by extensive simulation and a data analysis examining social interactions of ants.

IP34: "The recent development in spatio-temporal modelling"

08:30 - 10:10 Saturday, 6th January, 2024
B106 - Room 106, Babel Building
Chairs Tingjin Chu
Organiser: Tingjin Chu

ShengLi Tzeng: Robust Stationarity Testing and Contiguous Segmentation for Spatial Data

Jaehong Jeong: Wind vector modeling using space-time cross-covariance models

Hsin-Cheng Huang: Fast Spatial Prediction for Nonstationary Processes with a Divide-and-Conquer Strategy

Guoqi Qian: Discovering optimally representative dynamical locations (ORDL) in big multivariate spatiotemporal data: a case study of precipitation in Australia from space to ground sensors

IP34.01 Robust Stationarity Testing and Contiguous Segmentation for Spatial Data

Your name

ShengLi Tzeng

Abstract

In geostatistics, the assumption of stationarity is commonly applied to the underlying processes of interest. While these assumptions remain valid within constrained spatial domains, their applicability diminishes when addressing large geographical domains or intricate spatial phenomena. We introduce an innovative methodology tailored for the evaluation of stationarity assumption frequently employed in geostatistics. We look at a particular kind of nonstationarity, where the spatial covariance varies across the spatial domain. Our method utilizes robust local estimates of spatial covariances to calculate a statistic representing neighboring spatial dependency and employs Voronoi tessellations to cluster data locations for hypothesis testing. Additionally, when the assumption of stationarity is violated, we present a systematic framework for discerning nonstationary patterns. This is achieved by dividing the region into several contiguous subregions with more homogeneous and close-to-stationary properties. The optimal number of partitions is determined using the Bayesian information criterion, ensuring the accuracy and efficiency of our method. It is worth noting that our proposed method is applicable to gridded data or irregularly spaced data, making it a versatile tool for exploring a wide range of spatial datasets.

IP34.02 Wind vector modeling using space-time cross-covariance models

Your name

Jaehong Jeong

Abstract

As the risk posed by climate change becomes more apparent, countries around the world are constantly looking for alternative energy sources. Wind energy has significant potential for future energy portfolios without negative environmental impacts. We analyze wind vectors in East Asia from the fifth-generation ECMWF atmospheric reanalysis. To model the wind vectors, we consider non-Gaussian processes based on the Tukey g-and-h transformation, along with multivariate covariance functions. The proposed model can account for non-Gaussian features and some nonstationary dependence structures of the wind vectors. In addition, a two-step inference scheme coupled with the composite likelihood method is applied to handle the computational issues posed by a large data set. We show that the proposed method with a covariance structure reflecting the nonstationarity due to the latitude and the land-ocean differences leads to better predictions of wind speed and wind potential, which is crucial for wind power generation planning.

* This is joint work with Won Chang (University of Cincinnati).

IP34.03 Fast Spatial Prediction for Nonstationary Processes with a Divide-and-Conquer Strategy

Hsin-Cheng Huang

Your name

Hsin-Cheng Huang

Abstract

Spatial data over a large domain generally shows nonstationary features. However, appropriately specifying a nonstationary covariance function is complex, and the computation of the corresponding inverse matrix in kriging is intractable when the dataset is massive. In this research, we propose a methodology based on a linear combination of stationary processes with spatially varying weights for nonstationary spatial modeling.

Unlike existing approaches that typically assume the stationary processes are uncorrelated, we model them as a multivariate spatial process defined over the entire domain. Our spatial model allows the spatial covariance function to vary smoothly or sharply across regions governed by a tuning parameter. Notably, it reduces to a global stationary process when all stationary components share a common spatial covariance structure. We develop a divide-and-conquer strategy for fast spatial prediction. Numerical results show it is effective in nonstationary spatial covariance function estimation and spatial prediction.

IP34.04 Discovering optimally representative dynamical locations (ORDL) in big multivariate spatiotemporal data: a case study of precipitation in Australia from space to ground sensors

Your name

Guoqi Qian

Abstract

We develop a method for discovering a set of optimally representative dynamical locations (ORDL), a small subset of observed locations that are the most informative of the dynamics of a real complex system, as embodied in big spatiotemporal data. We achieve this through a two-pronged approach: (a) by reducing the multivariate time series data into a small set of time series with minimal loss of information on the dynamics of the system, (b) by exploiting the best that remote sensing and in-situ observations can offer. In the former, we extend the recently proposed empirical dynamical quantiles for univariate time series to multivariate data using a directional statistical depth measure and principal eigen-decomposition method. In the latter, we perform data fusion to leverage remotely sensed precipitation from multiple satellite platforms in addition to ground-based rain gauges to improve overall accuracy and spatial coverage. We demonstrate our method in the context of precipitation data over 2003-2021 for Australia. Of the six states, the location, ranking and number of ORDL suggest that Queensland has seen the most significant variability in precipitation while that in Victoria has remained relatively stable. Finally, this study has uncovered ungauged locations in data-sparse regions of Australia where the installation of future rain gauges can optimally represent precipitation dynamics in the region under a changing climate. (Joint work with Benjamin Hines and Antoinette Tordesillas)

IP62: "New development of statistical learning in data science"

08:30 - 10:10 Saturday, 6th January, 2024
OA239 - Room 239, Old Arts Building
Chairs Yuehan Yang
Organiser: Yuehan Yang

Yanyan Liu: Sparse Learning Via a Novel Penalty and a Fast Solver

Rui Pan: Large-scale Multi-layer Academic Networks Derived from Statistical Publications

Jialiang Li: Autoregressive Networks

Yuehan Yang: Randomization-based Joint Central Limit Theorem and Efficient Covariate Adjustment in Randomized Block 2K Factorial Experiment

---

IP62.01 Sparse LearningVia a Novel Penalty and a Fast Solver

Your name

Yanyan Liu

Abstract

In machine learning and statistics, the penalized regression methods are the main tools for variable selection (or feature selection) in high-dimensional sparse data analysis. Due to the nonsmoothness of the associated thresholding operators of commonly used penalties such as the least absolute shrinkage and selection operator (LASSO), the smoothly clipped absolute deviation (SCAD), and the minimax concave penalty (MCP), the classical Newton–Raphson algorithm cannot be used. In this article, we propose a cubic Hermite interpolation penalty (CHIP) with a smoothing thresholding operator. Theoretically, we establish the nonasymptotic estimation error bounds for the global minimizer of the CHIP penalized high-dimensional linear regression. Moreover, we show that the estimated support coincides with the target support with a high probability. We derive the Karush–Kuhn–Tucker (KKT) condition for the CHIP penalized estimator and then develop a support detection-based Newton–Raphson (SDNR) algorithm to solve it. Simulation studies demonstrate that the proposed method performs well in a wide range of finite sample situations. We also illustrate the application of our method with a real data example.

IP62.02 Large-scale Multi-layer Academic Networks Derived from Statistical Publications

Your name

Rui Pan

Abstract

The utilization of multi-layer network structures now enables the explanation of complex systems in nature from multiple perspectives. Multi-layer academic networks capture diverse relationships among academic entities, facilitating the study of academic development and the prediction of future directions. However, there are currently few academic network datasets that simultaneously consider multi-layer academic networks; often, they only include a single layer. In this study, we provide a large-scale multi-layer academic network dataset, namely, LMANStat, which includes collaboration, co-institution, citation, co-citation, journal citation, author citation, author-paper and keyword co-occurrence networks. Furthermore, each layer of the multi-layer academic network is dynamic. Additionally, we expand the attributes of nodes, such as authors' research interests, productivity, region and institution. Supported by this dataset, it is possible to study the development and evolution of statistical disciplines from multiple perspectives. This dataset also provides fertile ground for studying complex systems with multi-layer structures.

IP62.03 Autoregressive Networks

Your name

Jialiang Li

Abstract

We propose a first-order autoregressive (i.e. AR(1)) model for dynamic network processes in which edges change over time while nodes remain unchanged. The model depicts the dynamic changes explicitly. It also facilitates simple and efficient statistical inference methods including a permutation test for diagnostic checking for the fitted network models. The proposed model can be applied to the network processes with various underlying structures but with independent edges. As an illustration, an AR(1) stochastic block model has been investigated in depth, which characterizes the latent communities by the transition probabilities over time. This leads to a new and more effective spectral clustering algorithm for identifying the latent communities. We have derived a finite sample condition under which the perfect recovery of the community structure can be achieved by the newly defined spectral clustering algorithm.

Furthermore the inference for a change point is incorporated into the AR(1) stochastic block model to cater for possible structure changes. We have derived the explicit error rates for the maximum likelihood estimator of the change-point. Application with three real data sets illustrates both relevance and usefulness of the proposed AR(1) models and the associate inference methods.

IP62.04 Randomization-based Joint Central Limit Theorem and Efficient Covariate Adjustment in Randomized Block 2K Factorial Experiments

Yuehan Yang

Central University of Finance and Economics, Beijing, China

Your name

Yuehan Yang

Abstract

Randomized block factorial experiments are widely used in industrial engineering, clinical trials, and social science. Researchers often use a linear model and analysis of covariance to analyze experimental results; however, limited studies have addressed the validity and robustness of the resulting inferences because assumptions for a linear model might not be justified by randomization in randomized block factorial experiments. In this paper, we establish a new finite population joint central limit theorem for usual (unadjusted) factorial effect estimators in randomized block 2^K factorial experiments. Our theorem is obtained under a randomization-based inference framework, making use of an extension of the vector form of the Wald--Wolfowitz--Hoeffding theorem for a linear rank statistic. It is robust to model misspecification, numbers of blocks, block sizes, and propensity scores across blocks. To improve the estimation and inference efficiency, we propose four covariate adjustment methods. We show that under mild conditions, the resulting covariate-adjusted factorial effect estimators are consistent, jointly asymptotically normal, and generally more efficient than the unadjusted estimator. In addition, we propose Neyman-type conservative estimators for the asymptotic covariances to facilitate valid inferences. Simulation studies and a clinical trial data analysis demonstrate the benefits of the covariate adjustment methods.

IP30: "Recent Advances in Approximate and Generalized Bayesian methods"

08:30 - 10:10 Saturday, 6th January, 2024
EEBrown - Brown Room, Electrical Engineering Building
Chairs David Frazier
Organiser: David Frazier

Imke Botha: Component-wise iterative ensemble Kalman inversion for static Bayesian models with unknown measurement error covariance.

Ruben Loaiza-Maya: Efficient variational approximations for state space models

Takuo Matsubara: Hamiltonian Dynamics of Bayesian Inference Formalised by Arc Hamiltonian Systems

Minh-ngoc Tran: Natural gradient Variational Bayes without matrix inversion

---

IP30.01 **Component-wise iterative ensemble Kalman inversion for static Bayesian models with unknown measurement error covariance.**

Your name

Imke Botha

Abstract

Ensemble Kalman inversion (EKI) can be used for approximate inference of Gaussian likelihood models, where the mean is expensive to evaluate and the covariance is known. In this context, EKI methods are a fast alternative to Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC), as they generally require much fewer likelihood evaluations. Due to their reliance on a known covariance however, the applicability of EKI is limited as the true noise is rarely known in practice. We develop the component-wise iterative EKI (CW-IEKI) approach, which allows elements of the covariance matrix to be inferred alongside the model parameters at negligible extra cost. This is joint work with Chris Drovandi, Matthew Adams, Dan Tran and Frederick Bennett.

IP30.02 Efficient variational approximations for state space models

Your name

Ruben Loaiza

Abstract

Variational Bayes methods are a scalable estimation approach for many complex state space models. However, existing methods exhibit a trade-off between accurate estimation and computational efficiency. This paper proposes a variational approximation that mitigates this trade-off. This approximation is based on importance densities that have been proposed in the context of efficient importance sampling. By directly conditioning on the observed data, the proposed method produces an accurate approximation to the exact posterior distribution. Because the steps required for its calibration are computationally efficient, the approach is faster than existing variational Bayes methods. The proposed method can be applied to any state space model that has a closed-form measurement density function and a state transition distribution that belongs to the exponential family of distributions. We illustrate the method in numerical experiments with stochastic volatility models and a macroeconomic empirical application using a high-dimensional state space model.

IP30.03 Hamiltonian Dynamics of Bayesian Inference Formalised by Arc Hamiltonian Systems

Your name

Takuo Matsubara

Abstract

In this talk, we reveal an implication of the belief updating mechanism of Bayesian inference from the prior to the posterior as an infinitesimal change of a measure in an infinite-dimensional Hamiltonian flow. To this end, we establish a novel class of Hamiltonian systems, called arc Hamiltonian systems, for saddle Hamiltonian functions over infinite-dimensional metric spaces. Arc Hamiltonian systems generate a flow that satisfies the law of conservation of energy everywhere in a metric space. They are governed by an extension of Hamilton's equation formulated based on (i) the framework of arc fields and (ii) an infinite-dimensional gradient, termed the arc gradient, of a Hamiltonian function. We present two Hamiltonian functions, called the cumulant generating functional and the centred cumulant generating functional, over a metric space of log-likelihoods and measures.

The former characterises the posterior as a part of the arc gradient that induces a flow of log-likelihoods and non-negative measures. The latter characterises the difference of the posterior and the prior as a part of the arc gradient that induces a flow of log-likelihoods and probability measures.

IP30.04 Natural gradient Variational Bayes without matrix inversion

Your name

Minh-Ngoc Tran

Abstract

We present an approach for efficiently approximating the inverse of Fisher information, a key component in variational Bayes inference. A notable aspect of our approach is the avoidance of analytically computing the Fisher information matrix and its inversion. Instead, we introduce an iterative procedure for generating a sequence of matrices that converge to the inverse of Fisher information. Our natural gradient variational Bayes algorithm without matrix inversion is provably convergent and achieves a convergence rate of order O(log T/T), with T the number of iterations. We also obtain a central limit theorem for the iterates. Our algorithm exhibits versatility, making it applicable across a diverse array of variational Bayes domains, including Gaussian approximation and normalizing flow Variational Bayes. We offer a range of numerical examples to demonstrate the efficiency and reliability of the proposed method.

IP59: "Statistical Learning in Nonstandard Situations"

08:30 - 10:10 Saturday, 6th January, 2024
OA107 - Room 107, Old Arts Building
Chairs Minwoo Chae
Organiser: Seung Jun Shin

Byungwon Kim: Compositional data analysis by the square-root transformation: Application to NBA USG% data

Jun Song: Advances in functional predictor selection with its nonasymptotic bounds

Kyongwon Kim: On Modeling after Dimension Reduction

Seung Jun Shin: Variable Selection in AUC-optimizing Classification

IP59.01 Compositional data analysis by the square-root transformation: Application to NBA USG% data

Your name

Byungwon Kim

Abstract

Compositional data refers to data where the sum of the values of the components is a constant, hence the sample space is defined as a simplex making it impossible to apply statistical methods developed in the usual Euclidean vector space. A natural approach to overcome this restriction is to consider an appropriate transformation which moves the sample space onto the Euclidean space, and log-ratio typed transformations, such as the additive log-ratio (ALR), the centered log-ratio(CLR) and the isometric log-ratio(ILR) transformations, have been mostly conducted. However, in a high-dimensional compositional data such as a microbiome data, these log-ratio typed transformations are not appropriately working due to the sparsity which means that a substantial number of components take exact values of zero. In this work, we mainly suggest an alternative transformation, that is the square-root transformation which moves the original sample space onto the directional space. We compare the square-root transformation with the log-ratio typed transformation by the simulation study and the real data example. In the real data example, we applied both types of transformations to the USG% data obtained from NBA, and used a density based clustering method, DBSCAN, to show the result.

IP59.02 Advances in functional predictor selection with its nonasymptotic bounds

Your name

Jun Song

Abstract

This research introduces an innovative approach to functional predictor selection and estimation in a scalar-on-function regression model. Specifically, we integrate functional adaptive group-lasso type penalization for addressing regression issues with multivariate functional data predictors. This novel penalty, tailored for infinite-dimensional functional data, eases the traditionally rigorous assumptions required for theoretical validation. Our study establishes the consistency of this approach and delves into its nonasymptotic characteristics in the context of finite samples. Comprehensive simulations and real data applications show the method's effectiveness and credibility.

IP59.03 On Modeling after Dimension Reduction

Your name

Kyongwon Kim

Abstract

Despite the explosive development of sufficient dimension reduction methods, there has been a limited discussion for conducting statistical inference after dimension reduction. The outcome of dimension reduction is a set of sufficient predictors, but this is not the end of data analysis, furnished with confidence intervals and \$p\$-values for statistical significance. Currently, the common practice is to apply sufficient predictors into subsequent modeling as if they were the true predictors, but this leads to overly optimistic results of statistical inference. In this paper, we provide practical forms to show how errors from dimension reduction affect the standard error of regression parameters in a two-step procedure. In particular, we choose ordinary least squares as a dimension reduction method and linear and logistic regression as a modeling tool. We further investigate how to improve statistical inference when the number of covariates is larger than the sample size using the seeded regression method. In the numerical study, we compare coverage probabilities and the length of the confidence interval between before and after considering the errors induced from the dimension reduction step. Finally, we apply our approach to three real datasets covering a variety of modeling cases.

IP59.04 Variable Selection in AUC-optimizing Classification

Your name

Seung Jun Shin

Abstract

Optimizing the receiver operating characteristics (ROC) curve is often desired in imbalanced classification. In this article, we propose a binary classifier that optimizes the area under the ROC curve (AUC) penalty and is penalized by the smoothly clipped absolute deviance (SCAD) penalty, referred to as the SCAD-AUC estimator, and study its properties thoroughly. We established that the SCAD-AUC estimator possesses the oracle property in high dimension, enabling us to propose a consistent BIC-type information criterion that greatly facilitates the tuning procedure. Both simulated and real data analyses demonstrate the promising performance of the proposed SCAD-penalized AUC-optimizing classifier in terms of variable selection and prediction.

CP10: Contributed Paper Session

08:30 - 10:10 Saturday, 6th January, 2024
OA224 - Room 224, Old Arts Building
Chairs Edmund Lau

Masanari Kimura: Equivalence of Geodesics and Importance Weighting from the Perspective of Information Geometry

Matthew Shen: Introducing GutGPT: An AI Chatbot to Provide Interpretable and Context-Guided Risk Assessment for Patients with Gastrointestinal Bleeding

Selina Drews: Analysis of the expected L2 error of an over-parametrized deep neural network estimate learned by gradient descent without regularization

Edmund Lau: Quantifying degeneracy in singular models via the learning coefficient

Mongju Jeong: Exploring Spatial Dynamics in Regression Coefficients: A Bayesian Regularization Method with Clustering

CP10.01 Equivalence of Geodesics and Importance Weighting from the Perspective of Information Geometry

Your name

Masanari Kimura

Abstract

In many machine learning settings, importance weighting plays a central role. For instance, under the covariate shift assumption, it is known that by adopting the density ratio between the training distribution and the test distribution as importance weights, consistent parameter estimation is achievable even under this assumption. To begin with, we introduce that procedures of importance weighting can be equated with the selection of geodesics on a Riemannian manifold created by probability distributions. Such a framework that consider the geometry on the manifold created by sets of probability distributions is called the information geometry, and it is known that highly valuable in gaining a deep understanding of algorithms in statistics and machine learning. We introduce, from the perspective of Information Geometry, a generalization of Importance Weighted Empirical Risk Minimization induced by the equivalence between weighting strategies and geometric objects. Furthermore, we demonstrate that the parameter search for geometrically generalized covariate shift adaptation can be efficiently achieved, and this generalization outperforms existing methods based on the results of numerical experiments.

CP10.02 Introducing GutGPT: An AI Chatbot to Provide Interpretable and Context-Guided Risk Assessment for Patients with Gastrointestinal Bleeding

Your name

Matthew Shen

Abstract

Gastrointestinal bleeding (GIB) is the most common gastrointestinal diagnosis requiring hospitalization in the United States. Using electronic health record (EHR) data from three hospitals in the Yale New Haven Health system, we experiment with several machine learning and deep learning estimators and embedding methods and develop a pipeline to preprocess heterogeneous EHR data within the first four hours of patient presentation to automatically calculate GIB risk scores that can be then made available to emergency department providers for clinical decision support regarding need for hospital-based care. The final machine learning model identified patients who required a hospital-based intervention with an AUC > 0.9 on an external test set, outperforming existing clinical risk scores. Finally, we introduce Gut-GPT, a large language model that we envision will act as a mediator between clinicians and our machine learning model. Clinicians can query Gut-GPT in real-time about patient management options according to current guidelines in addition to a patient's model-predicted risk stratification along with relevant explainability metrics and plots of the prediction.

CP10.03 Analysis of the expected L_2 error of an over-parametrized deep neural network estimate learned by gradient descent without regularization

Your name

Selina Drews

Abstract

In this talk, we consider the estimation of a regression function from independent and identically distributed data. We introduce an estimate based on a deep neural network that is over-parametrized and trained by gradient descent to minimize the empirical L_2 risk, without any regularization. The neural network is over-parametrized in the sense that the number of parameters significantly exceeds the sample size.

Under appropriate conditions on the step size, the number of gradient descent steps, and a suitable random initialization of the parameters, we show that the estimate is universally consistent. Moreover, under the additional assumption that the regression function is Hölder smooth with Hölder exponent $p \in [1/2,1]$, a rate of convergence of approximately $n^{-1/(1+d)}$ of the expected L_2 error is derived.

CP10.04 Quantifying degeneracy in singular models via the learning coefficient

Your name

Edmund Lau

Abstract

Deep neural networks (DNN) are singular statistical models which exhibit complex degeneracies. In this work, we illustrate how a quantity known as the learning coefficient introduced in singular learning theory quantifies precisely the degree of degeneracy in deep neural networks. Importantly, we will demonstrate that degeneracy in DNN cannot be accounted for by simply counting the number of "flat" directions. We propose a computationally scalable approximation of a localized version of the learning coefficient using stochastic gradient Langevin dynamics. To validate our approach, we demonstrate its accuracy in low-dimensional models with known theoretical values. Importantly, the local learning coefficient can correctly recover the ordering of degeneracy between various parameter regions of interest. An experiment on MNIST shows the local learning coefficient can reveal the inductive bias of stochastic opitmizers for more or less degenerate critical points.

CP10.05 Exploring Spatial Dynamics in Regression Coefficients: A Bayesian Regularization Method with Clustering

Your name

Mongju Jeong

Abstract

Many fields are seeing expansive spatial datasets with increasing observations and covariates on a large spatial domain. As study domains expand, simple spatial assumptions of homogeneity in relationships might not capture complex dynamics. Using Bayesian partitioning with Voronoi tessellations, we address cluster-wise variations and effect changes across boundaries. We propose a Bayesian regularized spatially clustered coefficient (BRSCC) regression model, which can identify key covariates affecting the response variable and their spatial cluster patterns. The joint execution is facilitated through a latent indicator variable that shapes the prior model for each regression coefficient and cluster arrangement. We've developed a reversible jump MCMC-based algorithm and tested the model's effectiveness through simulation studies.

CP19: Contributed Paper Session

08:30 - 10:10 Saturday, 6th January, 2024
BG03 - Room G03, Babel Building
Chairs Dongming Huang

Hisayuki Hara: FAVAR Model with Instantaneous Effects and Its Identifiability

LuYi Shen: Online Hybrid Neural Network for Stock Price Prediction: A Case Study of High-Frequency Stock Trading in the Chinese Market.

Dongming Huang: Sliced Inverse Regression with Large Structural Dimensions

Shuya Nagayasu: Generalization Error of Bayesian Deep Neural Network with non analytic activation function

Seeun Park: Combined quantile forecasting for high-dimensional non-Gaussian data

---

CP19.01 FAVAR Model with Instantaneous Effects and Its Identifiability

Your name

Hisayuki Hara

Abstract

In recent years, modeling using latent variables has attracted attention in time series data analysis. FAVAR model is a model that combines a dynamic factor model with a factor analysis-type VAR model. The FAVAR model is a VAR-type model using time-lagged variables. For data like annual data, models with instantaneous causal effects are often preferred. In this presentation, we propose the FAVAR LiNGAM model, which combines FAVAR model and VAR LiNGAM model, as a model that considers the instantaneous causal effect. We give a sufficient condition for the FAVAR LiNGAM model to be identifiable and provide an estimation algorithm. Computer experiments showing the validity of the proposed method will also be presented.

CP19.02 Online Hybrid Neural Network for Stock Price Prediction: A Case Study of High-Frequency Stock Trading in the Chinese Market.

Your name

Luyi Shen

Abstract

This study introduces the Online LGT (O-LGT) for high-frequency trading (HFT) forecasts, efficiently predicting time-series data characterized by low signal-to-noise ratio, non-stationarity, and non-linearity. O-LGT, integrating LSTM, GRU, and Transformer, facilitates fast, accurate predictions crucial for exploiting HFT price discrepancies. The innovative method efficiently manages storage for ultra-fast computing, using calculated outputs (instead of previous trading data) for immediate future forecasts, thus speeding up computations. Performance evaluations using high-frequency Limit Order Book (LOB) data reveal O-LGT's comparable speed to conventional HFT models with enhanced accuracy. Despite a slight accuracy trade-off, O-LGT is 12 to 64 times faster than other high-accuracy LOB data models in the Chinese market.

CP19.03 Sliced Inverse Regression with Large Structural Dimensions

Your name

Dongming Huang

Abstract

The central space of a joint distribution (X, Y) is the minimal subspace S such that $Y \perp X \mid P X$ where P is the projection onto S. Sliced inverse regression (SIR), one of the most popular methods for estimating the central space, often performs poorly when the structural dimension $d = \dim(S)$ is large (e.g., ≥ 5). In this paper, we demonstrate that the generalized signal- noise-ratio (gSNR) tends to be extremely small for a general multiple-index model when d is large. Then we determine the minimax rate for estimating the central space over a large class of high dimensional distributions with a large structural dimension d (i.e., there is no constant upper bound on d) in the low gSNR regime. This result not only extends the existing minimax rate results for estimating the central space of distributions with fixed d to that with a large d, but also clarifies that the degradation in SIR performance is caused by the decay of signal strength. The technical tools developed here might be of independent interest for studying other central space estimation methods.

CP19.04 Generalization Error of Bayesian Deep Neural Network with non analytic activation function

Your name

Shuya Nagayasu

Abstract

Deep Neural Network is widely used in practical way. However why large scale neural network can generalize have not been completely explained yet. In Bayesian Learning, the generalization error of non-identifiable learning model including neural networks has been revealed by using algebraic geometrical methods. However, analytic activation function was needed for applying such methods.

In this presentation, we introduce recent studies about the generalization error of Bayesian Deep Neural Network with relu activation function which is non-analytic. These studies revealed that in Bayesian Deep Neural Networks with relu activation function, the mass of Bayesian Posterior is concentrated to the neighborhood of optimal parameter such as convex at only smallest dimension for approximating true and flat at other dimensions. This property come from non-analyticity makes the generalization error smaller. The theoretical value of the generalization error does not depend on the number of dimensions of the model. This result is expected to explain the mechanism of overparametrized neural network generalizing.

CP19.05 Combined quantile forecasting for high-dimensional non-Gaussian data

Your name

Seeun Park

Abstract

This study proposes a novel method for forecasting a scalar variable based on high-dimensional predictors that is applicable to various data distributions. In the literature, one of the popular approaches for forecasting with many predictors is to use factor models. However, these traditional methods are ineffective when the data exhibit non-Gaussian characteristics such as skewness or heavy tails. In this study, we newly utilize a quantile factor model to extract quantile factors that describe specific quantiles of the data beyond the mean factor. We then build a quantile-based forecast model using the estimated quantile factors at different quantile levels as predictors.

Finally, the predicted values at the various quantile levels are combined into a single forecast as a weighted average with weights determined by a Markov chain based on past trends of the target variable. The main idea of the proposed method is to incorporate a quantile approach to a forecasting method to handle non-Gaussian characteristics effectively. The performance of the proposed method is evaluated through a simulation study and real data analysis of $PM_{2.5}$ data in South Korea, where the proposed method outperforms other existing methods in most cases.

DL13: Distinguished Lecture Session

10:50 - 12:30 Saturday, 6th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Ming-Yen Cheng

Lixing Zhu: Change point detection for tensors with heterogeneous slices

Zhou Yu: Deep Nonlinear Sufficient Dimension Reduction

Dong Xia: Multiple Testing of Linear Forms for Noisy Matrix Completion

DL13.01 Change point detection for tensors with heterogeneous slices

Your name

Lixing Zhu

Abstract

In many applications, tensor data may consist of heterogeneous slices according to a categorical mode, and independent but not identically distributed error tensors over time. To detect change structures in such tensor data, we define a mode-based signal-screening Frobenius distance for the moving sums of slices to handle both dense and sparse model structures of the tensors. Based on this distance, we construct a mode-based signal statistic using a sequence of ratios with mode-based adaptive-to-change ridge functions. The number of changes and their locations can be consistently estimated in certain senses, and the confidence intervals of the locations of change points are constructed when the standardized error tensors are homogenous. The results hold when the size of the tensor and the number of change points diverge at certain rates, respectively. Numerical studies are conducted to examine the finite sample performances of the proposed method. We also analyze two real data examples for illustration.

DL13.02 Deep Nonlinear Sufficient Dimension Reduction

Your name

Zhou Yu

Abstract

Linear sufficient dimension reduction, as exemplified by sliced inverse regression, has seen substantial development in the past thirty years. However, with the advent of more complex scenarios, nonlinear dimension reduction has become a more general topic that gains considerable interest recently. We introduce a novel method for nonlinear sufficient dimension reduction, utilizing the generalized martingale difference divergence measure in conjunction with deep neural networks. And two optimization schemes considered, based on the fascinating deep neural networks, exhibit higher efficiency and flexibility compared to the classical eigende composition of linear operators. Moreover, we systematically investigate the slow rate and fast rate for the estimation error based on advanced U-process theory. Remarkably, the fast rate is nearly minimax optimal. The effectiveness of the deep nonlinear sufficient dimension reduction methods is demonstrated through simulations and real data analysis.

DL13.03 Multiple Testing of Linear Forms for Noisy Matrix Completion

Your name

Dong Xia

Abstract

Many important tasks of large-scale recommender systems can be naturally cast as testing multiple linear forms for noisy matrix completion. These problems, however, present unique challenges because of the subtle bias-and-variance tradeoff of and an intricate dependence among the estimated entries induced by the low-rank structure. In this paper, we develop a general approach to overcome these difficulties by introducing new statistics for individual tests with sharp asymptotics both marginally and jointly, and utilizing them to control the false discovery rate (FDR) via a data splitting and symmetric aggregation scheme. We show that valid FDR control can be achieved with guaranteed power under nearly optimal sample size requirement. Extensive numerical simulations and real data examples underscore the merits of the proposed method.

IP58: "Statistical Inference of Network and Dependent Data"

10:50 - 12:30 Saturday, 6th January, 2024
B305 - Room 305, Babel Building
Chairs Somabha Mukherjee
Organiser: Somabha Mukherjee

Jing Lei: Recent Advances in Tensor and Dynamic Stochastic Block Models

Anderson Zhang: Fundamental Limits of Spectral Clustering in Stochastic Block Models

Bhaswar Bhattacharya: Higher-Order Graphon Theory: Fluctuations and Inference

IP58.01 Recent Advances in Tensor and Dynamic Stochastic Block Models

Your name

Jing Lei

Abstract

We consider the problem of estimating common community structures in multi-layer stochastic block models, where each single layer may not have sufficient signal strength to recover the full community structure. We analyze the effect of having a large number of layers on optimal estimation threshold. Interestingly, the number of layers has a linear effect on the information-theoretic estimation threshold, and the effect becomes square-root if we restrict to polynomial-time algorithms assuming a low-degree polynomial hardness conjecture.

IP58.02 Fundamental Limits of Spectral Clustering in Stochastic Block Models

Your name

Anderson Ye Zhang

Abstract

We give a precise characterization of the performance of spectral clustering for community detection under Stochastic Block Models by carrying out sharp statistical analysis. We show spectral clustering has an exponentially small error with matching upper and lower bounds that have the same exponent, including the sharp leading constant.

The fundamental limits established for the spectral clustering hold for networks with multiple and imbalanced communities and sparse networks with degrees far smaller than $\log n$. The key to our results is a novel truncated ℓ_2 perturbation analysis for eigenvectors and a new analysis idea of eigenvectors truncation.

IP58.03 Higher-Order Graphon Theory: Fluctuations and Inference

Your name

Bhaswar B. Bhattacharya

Abstract

Motifs (patterns of subgraphs), such as edges and triangles, encode important structural information about the geometry of a network. Consequently, counting motifs in a large network is an important statistical and computational problem. In this talk we will consider the problem of estimating motif densities and fluctuations of subgraph counts in an inhomogeneous random graph sampled from a graphon. We will show that the limiting distributions of subgraph counts can be Gaussian or non-Gaussian, depending on a notion of regularity of subgraphs with respect to the graphon. Using these results and a novel multiplier bootstrap for graphons, we will construct joint confidence sets for the motif densities. Applications to various network hypothesis testing problems will be discussed.

(Joint work with Anirban Chatterjee, Svante Janson, and Soham Dan)

IP21: "Recent Advances on Functional and Complex Data Analysis"

10:50 - 12:30 Saturday, 6th January, 2024
EEBrown - Brown Room, Electrical Engineering Building
Chairs Zhenhua Lin
Organiser: Zhenhua Lin

Lijia Wang: Local perspectives in latent space social networks

Zhigang Yao: Manifold Fitting

Haozhen Shu: Simultaneous Inference for functional data by bootstrap

---

IP21.02 Local perspectives in latent space social networks

Your name

Lijia Wang

Abstract

Statistical network research frequently estimates network graph features using data from the complete network or information pooled from multiple subgraphs. However, in real-world social networks, decision-making often hinges on individuals' localized perspectives. Noting this divergence, Han et al. (2022) introduced an individual-centered partial information framework to depict individuals' local network views, thereby aiding the understanding of individual local perspectives of the network. Existing research within this framework focuses on community detection and heavily depends on community structures, as it uses stochastic block models. This approach might not fully encapsulate the complex nature of social network structures, as social actors seldom conform to simplistic patterns, such as clusters. Aware of this complexity, our research seeks to delve into individual-centered partial information within broader social network models, specifically latent space models. The flexibility, strong interpretability, and visualization capabilities of latent space models will help foster a more nuanced understanding of local perspectives within social networks.

IP21.03 Manifold Fitting

Your name

Zhigang Yao

Abstract

This manifold fitting problem can go back to H. Whitney's work in the early 1930s (Whitney (1992)), and finally has been answered in recent years by C. Fefferman's works (Fefferman, 2006, 2005). The solution to the Whitney extension problem leads to new insights for data interpolation and inspires the formulation of the Geometric Whitney Problems (Fefferman et al. (2020, 2021a)): Assume that we are given a set $Y \subset \mathbb{R}^D$. When can we construct a smooth d-dimensional submanifold $\widehat{M} \subset \mathbb{R}^D$ to approximate Y, and how well can $\widehat{M}$ estimate Y in terms of distance and smoothness? To address these problems, various mathematical approaches have been proposed (see Fefferman et al. (2016, 2018, 2021b)). However, many of these methods rely on restrictive assumptions, making extending them to efficient and workable algorithms challenging. As the manifold hypothesis (non-Euclidean structure exploration) continues to be a foundational element in statistics, the manifold fitting Problem, merits further exploration and discussion within the modern statistical community.

IP21.04 Simultaneous Inference for functional data by bootstrap.

Your name

Haozhen Shu

Abstract

In this paper, we introduce a novel approach for constructing a simultaneous confidence band (SCB) for multi-dimensional functional parameters using local polynomial estimators. Our focus is on the Gaussian and bootstrap approximation of the studentized process, which allows us to derive the validity of the SCB. Our proposed method is applicable to both sparse and dense functional data, and we also highlight the phase transition phenomenon.

IP02: "Gradient Descent and its Statistical Theory"

10:50 - 12:30 Saturday, 6th January, 2024
AW153 - Room 153, Arts West North Building
Chairs Yingqiu Zhu
Organiser: Hansheng Wang

Yingqiu Zhu: Automatic, Dynamic, and Nearly Optimal Learning Rate Specification via Local Quadratic Approximation

Haobo Qi: Statistical Analysis of Fixed Mini-Batch Gradient Descent Estimator

Yuan Gao: An Asymptotic Analysis of Random Partition Based Minibatch Momentum Methods for Linear Regression Models

Shuyuan Wu: Network Gradient Descent Algorithm for Decentralized Federated Learning

---

IP02.01 Automatic, Dynamic, and Nearly Optimal Learning Rate Specification via Local Quadratic Approximation

Your name

Yingqiu Zhu

Abstract

In deep learning tasks, the update step size determined by the learning rate at each iteration plays a critical role in gradient-based optimization. However, determining the appropriate learning rate in practice typically relies on subjective judgement. In this work, we propose a novel optimization method based on local quadratic approximation (LQA). In each update step, given the gradient direction, we locally approximate the loss function by using a standard quadratic function of the learning rate. Subsequently, we propose an approximation step to obtain a nearly optimal learning rate in a computationally efficient manner. The proposed LQA method has three important features. First, the learning rate is automatically determined in each update step. Second, it is dynamically adjusted according to the current loss function value and the parameter estimates. Third, with the gradient direction fixed, the proposed method attains a nearly maximum reduction in the loss function. Extensive experiments were conducted to prove the effectiveness of the proposed LQA method.

IP02.02 Statistical Analysis of Fixed Mini-Batch Gradient Descent Estimator

Your name

Haobo Qi

Abstract

We study here a fixed mini-batch gradient descent (FMGD) algorithm to solve optimization problems with massive datasets. In FMGD, the whole sample is split into multiple nonoverlapping partitions. Once the partitions are formed, they are then fixed throughout the rest of the algorithm. For convenience, we refer to the fixed partitions as fixed mini-batches. Then for each computation iteration, the gradients are sequentially calculated on each fixed mini-batch. Because the size of fixed mini-batches is typically much smaller than the whole sample size, it can be easily computed. This leads to much reduced computation cost for each computational iteration. It makes FMGD computationally efficient and practically more feasible. To demonstrate the theoretical properties of FMGD, we start with a linear regression model with a constant learning rate. We study its numerical convergence and statistical efficiency properties. We find that sufficiently small learning rates are necessarily required for both numerical convergence and statistical efficiency. Nevertheless, an extremely small learning rate might lead to painfully slow numerical convergence. To solve the problem, a diminishing learning rate scheduling strategy can be used. This leads to the FMGD estimator with faster numerical convergence and better statistical efficiency. Finally, the FMGD algorithms with random shuffling and a general loss function are also studied. Supplementary materials for this article are available online.

IP02.03 An Asymptotic Analysis of Random Partition Based Minibatch Momentum Methods for Linear Regression Models

Your name

Yuan Gao

Abstract

Momentum methods have been shown to accelerate the convergence of the standard gradient descent algorithm in practice and theory. In particular, the random partition based minibatch gradient descent methods with momentum (MGDM) are widely used to solve large-scale optimization problems with massive datasets. Despite the great popularity of the MGDM methods in practice, their theoretical properties are still underexplored.

To this end, we investigate the theoretical properties of MGDM methods based on the linear regression models. We first study the numerical convergence properties of the MGDM algorithm and derive the conditions for faster numerical convergence rate. In addition, we explore the relationship between the statistical properties of the resulting MGDM estimator and the tuning parameters. Based on these theoretical findings, we give the conditions for the resulting estimator to achieve the optimal statistical efficiency. Finally, extensive numerical experiments are conducted to verify our theoretical results.

IP02.04 Network Gradient Descent Algorithm for Decentralized Federated Learning

Your name

Shuyuan Wu

Abstract

We study a fully decentralized federated learning algorithm, which is a novel gradient descent algorithm executed on a communication-based network. For convenience, we refer to it as a network gradient descent (NGD) method. In the NGD method, only statistics (e.g., parameter estimates) need to be communicated, minimizing the risk of privacy. Meanwhile, different clients communicate with each other directly according to a carefully designed network structure without a central master. This greatly enhances the reliability of the entire algorithm. Those nice properties inspire us to carefully study the NGD method both theoretically and numerically. Theoretically, we start with a classical linear regression model. We find that both the learning rate and the network structure play significant roles in determining the NGD estimator's statistical efficiency. The resulting NGD estimator can be statistically as efficient as the global estimator, if the learning rate is sufficiently small and the network structure is well balanced, even if the data are distributed heterogeneously. Those interesting findings are then extended to general models and loss functions. Extensive numerical studies are presented to corroborate our theoretical findings. Classical deep learning models are also presented for illustration purpose.

IP11: "Point processes: asymptotics and applications"

10:50 - 12:30 Saturday, 6th January, 2024
B106 - Room 106, Babel Building
Chairs Aihua Xia
Organiser: Aihua Xia

J. E. Yukich: Limit theory for statistics of dynamic spatial random models

Leoni Carla Wirth: Metric-based methods for point processes and random graphs

Gopalan Nair: Analysing point patterns on a linear networks

Nicolas Privault: Moments of point processes

IP11.01 Limit theory for statistics of dynamic spatial random models

Your name

Joseph Yukich

Abstract

Abstract: We establish the limit theory for statistics of spatial random models evolving over a time domain and which are asymptotically de-correlated over spatial domains. The three sources of model randomness given by the random collection of particle locations, their random initial states, and the system evolution give rise to point processes with interacting time-evolving marks. The model randomness may be controlled via weak geometric conditions on the input and system evolution, giving the limit theory for statistics of the model as the spatial domain increases up to R^d . This gives weak laws of large numbers, variance asymptotics, and asymptotic normality for statistics of continuum versions of spin models, interacting diffusion models, and interacting particle systems. The talk is based on joint work with B. Blaszczyszyn and D. Yogeshwaran.

IP11.02 Metric-based methods for point processes and random graphs

Your name

Leoni Carla Wirth

Abstract

In the field of spatial point processes, the OSPA metric and the TT metric are two intuitive metrics between point patterns of (possibly) different sizes that are based on optimal assignments. We discuss recent research showing their suitability for statistical analysis and how they may be employed to investigate weak convergence of point processes via Stein's method.

Building upon these two point pattern metrics, we introduce a metric between spatial graphs of (possibly) different sizes that, additional to vertex differences, includes the discrepancy of edge structures in the optimal assignment. We present a statistical application to demonstrate its beneficial nature and introduce a representation of a graph as a pair of point processes to derive a new type of (random) graph convergence. Finally, we study this random graph convergence by developing Stein's method and deriving convergence rates.

IP11.03 Analysing point patterns on a linear networks

Your name

Gopalan Nair

Abstract

Recent years, a rapid increase in availability of point pattern data on linear network, in many applications areas, has spurred enormous interest in the development of statistical methodologies. In this talk we review some of these interesting developments. Due to geometrical complexities of network, methodological problems and risk of errors and the lack of homogeneity, methods for analysis of such data is challenging and significantly different to the classical methodology of spatial statistics based on stationary processes. Here we present those developments and methods, illustrating them with datasets from few different applications.

The statistical methodologies developed include, diffusion and other kernel density estimations and their bandwidth selection, estimation of relative risk, parametric and non-parametric model estimation for intensity, estimation of K-function and correlation function, and point process models under different metrics. The *spatstat* package is used as the computational tool in the analysis.

IP11.04 **Moments of point processes**

Your name

Nicolas Privault

Abstract

We present a range of algorithms for the closed-form computation of moments of all orders of point processes such as Poisson, compound Poisson, Markovian jump, Hawkes, and Boolean-Poisson processes. Applications to density estimation, random graph connectivity, and convergence rates will be reviewed. Implementations in Maple, Mathematica and Sage are available.

IP17: "Statistical inference for stochastic processes and YUIMA package"

10:50 - 12:30 Saturday, 6th January, 2024
OA239 - Room 239, Old Arts Building
Chairs Kengo Kamatani
Organiser: Kengo Kamatani

Nakahiro Yoshida: Asymptotic expansion for batched bandits

Masayuki Uchida: Parameter estimation for discretely observed linear parabolic SPDEs in two space dimensions with small noises

Hiroki Masuda: Asymptotics and computation for robust Gaussian quasi-likelihood inference

Emanuele Guidotti: Asymptotic expansion formulas for diffusion processes based on the perturbation method

IP17.01 Asymptotic expansion for batched bandits

Your name

Nakahiro Yoshida

Abstract

We discuss an application of the asymptotic expansion theory to batched bandits. In the batched bandit, the environment of each stage is randomly set according to the random outcomes of the previous stage. We introduce a backwards asymptotic expansion formula and assess the backward propagation of errors. The asymptotic expansion method based on partial mixing was formulated in Yoshida (PTRF2004) and originally applied to asymptotic expansion of the additive functional of a partially mixing epsilon-Markov process such as a jump-diffusion process in the random environment. This is a joint work with Yechan Park.

IP17.02 Parameter estimation for discretely observed linear parabolic SPDEs in two space dimensions with small noises

Your name

Masayuki Uchida

Abstract

We investigate estimation methods for unknown parameters of linear parabolic second-order stochastic partial differential equations (SPDEs) in two spatial dimensions with small dispersion parameters driven by Q-Wiener processes using high-frequency data in time and space. Kaino and Uchida (2021,JJSD) studied parametric estimation for unknown parameters of a linear parabolic second order SPDE in one space dimension with a small dispersion parameter driven by the cylindrical Wiener process based on high-frequency data in time and space. We first present minimum contrast estimators for the three coefficient parameters of the SPDEs utilizing thinned data with respect to space. Next, we approximate the coordinate processes of the SPDEs using the previously mentioned minimum contrast estimators. Note that these coordinate processes are the Ornstein-Uhlenbeck processes with small dispersion parameters. Finally, we use the approximate coordinate processes to obtain parametric adaptive estimators for the remaining unknown parameters of the SPDEs. We also give an illustrative example and perform numerical simulations of the proposed estimators. This research is a collaborative effort with Yozo Tonaki (Osaka University) and Yusuke Kaino (Kobe University).

IP17.03 Asymptotics and computation for robust Gaussian quasi-likelihood inference

Your name

Hiroki Masuda

Abstract

The conventional Gaussian quasi-likelihood analysis for stochastic process models can be systematically robustified by perturbing the Kullback-Leibler divergence. Theoretical properties of the proposed estimator and its implementation in YUIMA package will be demonstrated.

IP17.04 Asymptotic expansion formulas for diffusion processes based on the perturbation method

Your name

Emanuele Guidotti

Abstract

Asymptotic expansion formulas for diffusion processes have been implemented in YUIMA. However, the asymptotic expansion scheme must be run whenever the initial conditions change because the general ODE system cannot be solved symbolically. In this talk, we discuss the possibility of reducing the general ODE system to a linear system for a particular choice of perturbation. In this case, the system can be solved symbolically so that all the coefficients and the final formulas depend symbolically on the initial conditions. Such implementation would provide accurate high-order approximations for the transition densities and moments of arbitrary diffusions that are fully symbolic.

IP10: "A High-Dimensional Multivariate Statistics and Their Advances"

10:50 - 12:30 Saturday, 6th January, 2024
OA107 - Room 107, Old Arts Building
Chairs Johan Lim
Organiser: Johan Lim

Yongho Jeon: Trace Ratio Optimization for High-Dimensional Multi-Class Discrimination

Long Feng: Sparse Kronecker product decomposition: a general framework of signal region detection in image regression

Gunwoong Park: Optimal Approach for Sub-Gaussian Linear Structural Equation Models

Seongoh Park: Testing correlation structure of matrix-variate data

IP10.01 Trace Ratio Optimization for High-Dimensional Multi-Class Discrimination

Your name

Yongho Jeon

Abstract

In multi-class discrimination with high-dimensional data, identifying a lower-dimensional subspace with maximum class separation is crucial. We propose a new optimization criterion for finding such a discriminant subspace, which is the ratio of two traces: the trace of between-class scatter matrix and the trace of within-class scatter matrix. Since this problem is not well-defined for high-dimensional data, we propose to regularize the within trace and maximize the between trace. A careful investigation reveals that this optimization has an innate connection to the eigenvalue decomposition of an indefinite matrix. For the sake of better interpretability of the classifier, we also consider a sparse estimation via a group-wise soft-thresholding. Interesting relationships between the proposed method and some classical methods such as Fisher's linear discriminant analysis and its variants are discussed. Empirical examples with simulated and real data sets suggest that the proposed method works well and is often better than some existing approaches in a wide range of problems, with respect to both variable selectivity and classification accuracy.

IP10.02 Sparse Kronecker product decomposition: a general framework of signal region detection in image regression

Your name

Long Feng

Abstract

This paper aims to present the first Frequentist framework on signal region detection in high-resolution and high-order image regression problems. Image data and scalar-on-image regression are intensively studied in recent years. However, most existing studies on such topics focussed on outcome prediction, while the research on region detection is rather limited, even though the latter is often more important. In this paper, we develop a general framework named Sparse Kronecker Product Decomposition (SKPD) to tackle this issue. The SKPD framework is general in the sense that it works for both matrices and tensors represented image data. Our framework includes one-term, multi-term, and nonlinear SKPDs. We propose nonconvex optimization problems for one-term and multi-term SKPDs and develop path-following algorithms for the nonconvex optimization. Under a Restricted Isometric Property, the computed solutions of the path-following algorithm are guaranteed to converge to the truth with a particularly chosen initialization even though the optimization is nonconvex. Moreover, the region detection consistency could also be guaranteed. The nonlinear SKPD is highly connected to shallow convolutional neural networks (CNN), particularly to CNN with one convolutional layer and one fully-connected layer. Effectiveness of SKPD is validated by real brain imaging data in the UK Biobank database.

IP10.03 Optimal Approach for Sub-Gaussian Linear Structural Equation Models

Your name

Gunwoong Park

Abstract

This study presents an improved identifiability condition for sub-Gaussian linear structural equation models (SEMs) that utilizes the maximum indegree of a graph, rather than the maximum degree of the moralized graph. Subsequently, it investigates the relationships between the proposed condition and the existing identifiability conditions. This study develops an optimal backward-learning approach for the proposed identifiable models.

The proposed algorithm consists of ordering and parent estimation steps in which both problems can be efficiently addressed using the best-subset-selection approach. Furthermore, the study provides theoretical guarantees for the proposed algorithm, where its sample complexity is optimal.

IP10.04 Testing correlation structure of matrix-variate data

Your name

Seongoh Park

Abstract

Separability (a Kronecker product) of a scatter matrix is one of favorable structures when multivariate heavy-tailed data are collected in a matrix form, due to its parsimonious representation. However, little attempt has been made to test separability beyond Gaussianity. In this paper, we present nonparametric separability tests that can be applied to a larger class of multivariate distributions not only including elliptical distributions but also generalized elliptical distributions and transelliptical distributions. The proposed test statistic exploits robustness of Tyler's M (or Kendall's tau) estimator and a likelihood function of a scaled variable. Since its distribution is hard to specify, we approximate the p-value using a permutation procedure, whose unbiasedness is obtained from the permutation invariance of multivariate paired data. Our simulation study demonstrates the efficacy of our method against other alternatives, and we apply it to rhesus monkey data and corpus callosum data.

IP53: "Recent developments in large-scale and massive data analysis"

10:50 - 12:30 Saturday, 6th January, 2024
BG03 - Room G03, Babel Building
Chairs Liuhua Peng
Organiser: Liuhua Peng

Haojie Ren: BASED: Bias-amended Semi-supervised Distribution Learning

Nan Zou: Statistical inference with optimal sampling: When does it work?

Pavel Krupskiy: On factor copula-based mixed regression models

Liuhua Peng: Distributed inference for two-sample U-statistics in massive data analysis

IP53.01 BASED: Bias-amended Semi-supervised Distribution Learning

Your name

Haojie Ren

Abstract

We consider the problem of constructing the distribution estimation in the semi-supervised setting. Differently from previous works on parameter inference, it is challenging to investigate semi-supervised distribution estimation due to the uniformity problem of a functional process. We proposed a novel and general framework to efficiently explore useful information in the unlabeled data via approximating one conditional distribution on covariates. The proposed semi-supervised distribution estimator is obtained from a K-fold cross-fitting strategy and achieves the consistency and asymptotic Gaussian process. It is also shown that the proposed estimator performs asymptotically more efficiently than the classical empirical cumulative distribution function (ECDF) under mild conditions. Further, we apply our method to characteristic inference, split conformal inference and heterogeneous treatment effects. Its advantages are demonstrated by extensive simulation and real data analysis.

IP53.02 Statistical inference with optimal sampling: When does it work?

Your name

Nan Zou

Abstract

In classic statistical inference, Ordinary Least Squares regression (OLS) has been the workhorse in studying the effect of one or more predictors on response. However, for datasets with massive sample sizes, which are increasingly prevalent these days, the OLS, with a computational complexity proportional to sample size times the square of the number of predictors, can be computationally infeasible.

To speed up the OLS for massive datasets, the optimal sampling OLS selects an optimal small subset of samples from the original massive number of samples. Despite its considerable popularity, little is known about the conditions required for optimal sampling OLS's asymptotical normality, which is critical in statistical inference.

This talk will first introduce the optimal sampling OLS procedure and then investigate the conditions for its asymptotical normality. Specifically, it seems (1) when the number of predictors is fixed, the optimal sampling OLS is asymptotically normal if and only if the OLS itself is asymptotically normal, and (2) when the number of predictors goes to infinity as the sample size goes to infinity, the optimal sampling OLS requires a more restrictive condition to be asymptotically normal than the OLS. This work is joint with A. Welsh.

IP53.03 On factor copula-based mixed regression models

Your name

Pavel Krupskiy

Abstract

In this article, a copula-based method for mixed regression models is proposed, where the conditional distribution of the response variable, given covariates, is modelled by a parametric family of continuous or discrete distributions, and the effect of a common latent variable of a cluster is modelled with a factor copula.

We show how to estimate the parameters of the copula and the parameters of the margins and find the asymptotic behaviour of the estimation errors. Numerical experiments are performed to assess the precision of the estimators for finite samples. An example of an application is given using COVID-19 vaccination hesitancy data from several countries. All developed methodologies are implemented in CopulaGAMM, available in CRAN.

IP53.04 Distributed inference for two-sample U-statistics in massive data analysis

Your name

Liuhua Peng

Abstract

This project considers distributed inference for two-sample U-statistics under the massive data setting. In order to reduce the computational complexity, this paper proposes distributed two-sample U-statistics and blockwise linear two-sample U-statistics. The blockwise linear two-sample U-statistic, which requires less communication cost, is more computationally efficient especially when the data are stored in different locations. The asymptotic properties of both types of distributed two-sample U-statistics are established. In addition, this paper proposes bootstrap algorithms to approximate the distributions of distributed two-sample U-statistics and blockwise linear two-sample U-statistics for both nondegenerate and degenerate cases. The distributed weighted bootstrap for the distributed two-sample U-statistic is new in the literature. The proposed bootstrap procedures are computationally efficient and are suitable for distributed computing platforms with theoretical guarantees. Extensive numerical studies illustrate that the proposed distributed approaches are feasible and effective.

CP01: Contributed Paper Session

10:50 - 12:30 Saturday, 6th January, 2024
OA224 - Room 224, Old Arts Building
Chairs Zhendong Huang

Wataru Urasaki: Geometric mean type of proportional reduction in variation measure for two-way contingency tables

Arijit Pyne: Robust and Efficient Estimation in Ordinal Response Models using the Density Power Divergence

Ha-Young Shin: Quantiles on global non-positive curvature spaces

Zhendong Huang: Estimating evolutionary and demographic parameters using succinct tree sequences

Seunghyeon Kim: Age-specific impacts of heat on mental health: An evidence from the capital city of South Korea

CP01.01 Geometric mean type of proportional reduction in variation measure for two-way contingency tables

Your name

Wataru Urasaki

Abstract

Various methods have been proposed for two-way contingency tables with nominal explanatory and objective variables to determine whether the two variables are independent. One of the methods is a test of independence, but it only determines whether there is independence. When the test shows that the variables are not independent or clearly related, one of the interests of the analyst is to quantitatively assess the degree of departure from independence, and thus the degree of association, at an interval of 0 to 1. Various association measures have been proposed, one of which is the proportional reduction in variation (PRV) measure which describes the PRV from the marginal distribution to the conditional distribution of the response. The conventional PRV measures can assess the association of the entire contingency table, but they cannot accurately assess the association for each explanatory variable.

In this study, we propose a geometric mean type of PRV (geoPRV) measure that aims to sensitively capture the association of each explanatory variable to the response variable by using a geometric mean, and it enables analysis without underestimation when there is partial association in cells of the contingency table. Furthermore, the geoPRV measure is constructed by using any functions that satisfy specific conditions, which has application advantages and makes it possible to express conventional PRV measures as geometric mean types in special cases. We also present some considerations from the results of the analysis with some artificial and actual data.

CP01.02 Robust and Efficient Estimation in Ordinal Response Models using the Density Power Divergence

Your name

Arijit Pyne

Abstract

In real life we frequently come across data sets that involve some independent explanatory variables, generating a set of ordinal responses. These ordinal responses may be thought to depend on a continuous latent variable and a set of unknown cut-off points. The latent variable is further assumed to be linearly related to the explanatory variables which in turn drive the ordinal responses. One way of estimating the unknown parameters is to find its MLE which is noted for being fully asymptotically efficient at the true model. However, a small proportion of outlying observations, e.g. responses incoherent to the categories or unbounded covariate(s) may destabilize the likelihood function to a great extent. Therefore, the reliability of the MLE is strongly affected. In this paper we use the density power divergence for parameter estimation such that the estimators are highly robust and asymptotically efficient. Asymptotic properties of the minimum density power divergence estimator are discussed. Its robustness is investigated through the influence function. Analytically, we have shown that the proposed estimator has a very high asymptotic breakdown point against data contamination. Numerically it is demonstrated that the proposed method yields slope estimates which never implode. In finite-sample simulation studies the proposed estimators are shown to outperform the MLE, producing more stable estimates at model misspecification. Also they are very competitive with other robust alternatives. The proposed estimators perform almost as good as the MLE at true model. Finally, we wrap up this article with an application on a real data example.

CP01.03 Quantiles on global non-positive curvature spaces

Your name

Ha-Young Shin

Abstract

This talk develops a notion of geometric quantiles on Hadamard spaces, also known as global non-positive curvature spaces. After giving some definitions and basic properties, including a necessary condition on the gradient of the quantile loss function at quantiles on Hadamard manifolds, we present asymptotic properties of sample quantiles on Hadamard manifolds such as strong consistency and joint asymptotic normality. Details on using a gradient descent algorithm to calculate quantiles, and in particular an explicit formula for the gradient of the quantile loss function, are provided for hyperbolic space, followed by experiments with both simulated and real single-cell RNA sequencing data.

CP01.04 Estimating evolutionary and demographic parameters using succinct tree sequences

Your name

Zhendong Huang

Abstract

The succinct tree sequence (TS) is an efficient data structure that approximates aspects of the evolutionary history underlying a sample of genome sequences. It has been widely used for efficient simulation and storage of genome datasets, but not yet as a basis for inference of demographic and evolutionary parameters, such as the mutation rate and population sizes. After demonstrating powerful inferences for evolutionary and demographic parameters based on the true TS, we develop TS-based approximate Bayesian computation for inferred TS, and show that much of the benefit from use of the TS is retained despite the inferred TS being locally inaccurate. Computational cost limits model complexity in TSABC, but we were able to incorporate unknown nuisance parameters and model misspecification, still finding big gains in parameter inference.

CP01.05 Age-specific impacts of heat on mental health: An evidence from the capital city of South Korea

Your name

Seunghyeon Kim

Abstract

Numerous studies have examined the impact of heat on mental health, with age being an important modifier in risk assessment. Age-stratified analysis are common, however, they fail to measure the effect size of different age groups. We aim to explore deeper insights into age-related vulnerability to heat using a more flexible approach. Instead of traditional age stratification, we used an extended distributed lag model, which provides a more flexible representation of age-related associations. The extended model, which follows the hypothesis that the impact of heat on mental health may be similar across neighborhood ages, reduced the dimensionality of the parameter space through a linear combination of predictive process basis functions and used neighborhood age information to improve the accuracy of parameter estimation. The model was adjusted for seasonality, long-term time trends, and age effects. We have identified specific age groups with increased susceptibility to elevated temperatures. For individuals aged 15-19 and 45-54, mental health admissions increased as the temperature on the day increased, whereas for individuals aged 20-44, the raised temperature decreased mental health admissions. Our findings highlight the importance of using flexible methods in environmental health research. Traditional age stratification, while useful, may overlook subtle age-related vulnerabilities. As climate change impacts become a reality, understanding these nuanced vulnerabilities will be critical for public health planning and interventions.

DL12: Distinguished Lecture Session

08:30 - 10:10 Sunday, 7th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Hans-Georg Mueller

Fang Yao: Theory of FPCA for discretized functional data

Zhenhua Lin: Statistical Inference for Functional Data via Bootstrapping

Ying Yang: Online Estimation for Functional Data

DL12.01 Theory of FPCA for discretized functional data

Your name

Fang Yao

Abstract

Functional data analysis is an important research field in statistics which treats data as random functions drawn from some infinite-dimensional functional space, and functional principal component analysis (FPCA) plays a central role for data reduction and representation. After nearly three decades of research, there remains a key problem unsolved, namely, the perturbation analysis of covariance operator for diverging number of eigencomponents obtained from noisy and discretely observed data. This is fundamental for studying models and methods based on FPCA, while there has not been much progress since the result obtained by Hall et al. (2006) for a fixed number of eigenfunction estimates. In this work, we establish a unified theory for this problem, deriving the moment bounds of eigenfunctions and asymptotic distributions of eigenvalues for a wide range of sampling schemes. We also exploit double truncation to derive the uniform convergence of such estimated eigenfunctions. The technical arguments in this work are useful for handling the perturbation series of discretely observed functional data and can be applied in models and methods involving inverse using FPCA as regularization, such as functional linear regression.

DL12.02 Statistical Inference for Functional Data via Bootstrapping

Your name

Zhenhua Lin

Abstract

Statistical inference is of fundamental importance and yet challenging in functional data analysis. In response to the challenge, a set of powerful bootstrap-based procedures are developed for hypothesis testing related to functional parameters, including the mean function and the slope function in functional linear regression. Validity and consistency of the proposed procedures are established, and convergence rates are derived. The proposed procedures are shown to enjoy excellent numeric performance especially when the sample size is limited while the signal is relatively weak.

DL12.03 Online Estimation for Functional Data

Your name

Ying Yang

Abstract

Functional data analysis has attracted considerable interest and is facing new challenges, one of which is the increasingly available data in a streaming manner. In this article we develop an online nonparametric method to dynamically update the estimates of mean and covariance functions for functional data. The kernel-type estimates can be decomposed into two sufficient statistics depending on the data-driven bandwidths. We propose to approximate the future optimal bandwidths by a sequence of dynamically changing candidates and combine the corresponding statistics across blocks to form the updated estimation. The proposed online method is easy to compute based on the stored sufficient statistics and the current data block. We derive the asymptotic normality and, more importantly, the relative efficiency lower bounds of the online estimates of mean and covariance functions. This provides insight into the relationship between estimation accuracy and computational cost driven by the length of candidate bandwidth sequence. Simulations and real data examples are provided to support such findings.

IP65: "Advances in Biostatistics"

08:30 - 10:10 Sunday, 7th January, 2024
AW153 - Room 153, Arts West North Building
Chairs Yuehan Yang
Organiser: Jialiang Li

Liming Xiang: Analysis of competing risks data with covariates subject to detection limits

Yan Shuo Tan: The Computational Curse of Big Data for Bayesian Additive Regression Trees: A Hitting Time Analysis

James Chun Yin Lee: Survival analysis with a random change-point

Yu Gu: A powerful transition model to assess treatment effect in COVID-19 clinical trials

IP65.01 Analysis of competing risks data with covariates subject to detection limits

Your name

Liming Xiang

Abstract

Competing risks occur commonly in survival analysis when subjects may experience multiple types of events and the occurrence of the primary event of interest can be precluded by a competing event. Challenges arise for the analysis of competing risks data with covariates subject to censoring due to detection limits. We propose a semiparametric multiple imputation method for inference under the subdistribution hazard model. Our proposed imputation model is compatible to the substantive model and can automatically utilize the information from the fully observed covariate values and the outcome data to efficiently impute the censored covariates based on rejection sampling. We establish consistency and asymptotic normality of the resulting estimator and demonstrate its promising finite sample performance via simulation studies. We illustrate its practical utility with the data from a study of community acquired pneumonia.

IP65.02 The Computational Curse of Big Data for Bayesian Additive Regression Trees: A Hitting Time Analysis

Your name

Yan Shuo Tan

Abstract

Bayesian Additive Regression Trees (BART) is a popular Bayesian non-parametric regression algorithm that is commonly used in causal inference. Its posterior is a distribution over sums of decision trees, and Markov Chain Monte Carlo (MCMC) is performed on this space to obtain approximate posterior samples. While the inferential properties of the BART posterior has been well-studied, there has been little theoretical work on the computational properties of BART. This is unfortunate because the BART sampler is notoriously often slow to mix. Prior work in this direction focused on mixing time lower bounds on tree structures, but these are unidentifiable parameters of the model. In this talk, we introduce a new method of quantifying the computational effectiveness of the BART sampler via hitting time lower bounds for the highest density posterior region, which conveniently captures all tree structures with the smallest bias and complexity. Across a range of different data generating models, we show theoretically that the hitting times grow with the training sample size, and we further illustrate this phenomenon with an extensive simulation study. Our results yield insights on why the BART sampler may experience computational issues and how to overcome these problems, such as via adjusting the temperature of the sampler.

IP65.03 Survival analysis with a random change-point

Your name

Chun Yin, Lee

Abstract

This research work is motivated by a breast cancer study, where the disease-free survival time of the patients is postulated to be regulated by the menopausal age, which is unobserved. As menopausal age varies across patients, a fixed change-point survival model may be inadequate. Therefore, we propose a novel proportional hazards model with a random change-point. We develop a nonparametric maximum likelihood estimation approach and devise a stable EM algorithm to compute the estimators.

Because the model is regular, we employ conventional likelihood theory for inference based on the asymptotic normality of the Euclidean parameter estimates, and the variance of the asymptotic distribution can be consistently estimated by a profile-likelihood approach. A simulation study demonstrates the satisfactory finite-sample performance of the proposed methods, which yield small bias and proper coverage probabilities. The methods are applied to the motivating breast cancer study.

IP65.04 A powerful transition model to assess treatment effect in COVID-19 clinical trials

Your name

Yu Gu

Abstract

In COVID-19 clinical trials of new treatments for hospitalized patients, clinical status is typically assessed on an ordinal scale and may suffer from missing data due to hospital discharge. We propose a novel transition model which addresses the missing data problem automatically and uses the entire course of clinical status to estimate the treatment effects. The estimators can be computed via a simple and stable EM algorithm and can be carried forward to a G-computation procedure to study various endpoints. We demonstrate through extensive simulation studies that the proposed estimation method performs well in realistic settings, and that the transition model + G-computation approach is more powerful and robust in detecting treatment effects than existing methods. Finally, we provide an application of our methods to the Adaptive COVID-19 Treatment Trial.

IP12: "Bayesian and Nonparametric Inference for Complex Data"

08:30 - 10:10 Sunday, 7th January, 2024
B305 - Room 305, Babel Building
Chairs Taeyoung Park
Organiser: Taeyoung Park

Minwoo Chae: Structured distribution estimation via generative adversarial networks

Chanmin Kim: A Bayesian Nonparametric Approach for Confounder Selection and Causal Estimation in High-Dimensional Observational Studies

Seongil Jo: Scalable and optimal Bayesian inference for sparse covariance matrices via screened beta-mixture prior

Cheng Li: Bayesian Fixed-Domain Asymptotics for Covariance Parameters in Spatial Gaussian Process Regression Models

---

IP12.01 Structured distribution estimation via generative adversarial networks

Your name

Minwoo Chae

Abstract

It is common in nonparametric estimation problems to impose a certain low-dimensional structure on the unknown parameter to avoid the curse of dimensionality. We consider a nonparametric distribution estimation problem with a novel structural assumption and investigate the use of generative adversarial networks (GAN) for estimating the unknown distribution. A convergence rate with respect to the L^1-Wasserstein metric is obtained, which depends only on the underlying structure and noise level. More interestingly, GAN achieves a faster convergence rate than the likelihood approach considered in the literature. We also discuss the minimax optimal rate with the proposed structural assumption.

IP12.02 A Bayesian Nonparametric Approach for Confounder Selection and Causal Estimation in High-Dimensional Observational Studies

Your name

CHANMIN KIM

Abstract

In observational studies, there is an increasing challenge of determining which covariates, potentially high-dimensional, are necessary to meet the assumption of ignorable treatment assignment for estimating causal effects. We introduce a Bayesian nonparametric method that accomplishes three main objectives: 1) prioritizes the inclusion of adjustment variables based on established confounder selection principles; 2) estimates causal effects while accommodating complex relationships among confounders, exposures, and outcomes; 3) provides causal estimates that consider uncertainty in confounder factors. Our approach relies on specifying multiple Bayesian additive regression tree models, linked together with a shared prior distribution. This prior distribution accumulates posterior selection probability for covariates based on their associations with both the exposure and the outcome of interest. Extensive simulation studies demonstrate the effectiveness of our method compared to similar approaches across various scenarios. We apply our method to investigate the causal impact of emissions from coal-fired power plants on ambient air pollution concentrations. The presence of potential confounding factors related to local and regional meteorological conditions introduces uncertainty concerning the role of a high-dimensional set of measured variables in confounding. Ultimately, our proposed method yields more efficient and consistent results across consecutive years compared to alternative methods, reinforcing the evidence of a causal link between SO2 emissions and ambient particulate pollution.

IP12.03 Scalable and optimal Bayesian inference for sparse covariance matrices via screened beta-mixture prior

Your name

Seongil Jo

Abstract

In this paper, we consider a high-dimensional setting where the number of variables p can grow to infinity as the sample size n gets larger. We assume that most of off-diagonal entries of the covariance matrix are zero. Several Bayesian methods for sparse covariance matrices have been proposed, but their computational speed is too slow, making them almost impossible to apply even to moderately high dimensions (e.g., $p \approx 200$). Motivated by this, we propose a scalable Bayesian method for large sparse covariance matrices. The main strategy of the proposed method is as follows: we first safely reduce the number of effective parameters in a covariance matrix, and then impose shrinkage priors only for selected nonzero off-diagonal entries. To this end, we suggest using the sure screening by keeping only the off-diagonal entries whose absolute sample correlation coefficients are larger than a threshold and furnishing the rests with zeros. It turns out that the proposed prior achieves the minimax or nearly minimax rate for sparse covariance matrices under the Frobenius norm. Therefore, it is not only computationally scalable but also optimal in terms of posterior convergence rate.

IP12.04 Bayesian Fixed-Domain Asymptotics for Covariance Parameters in Spatial Gaussian Process Regression Models

Your name

Cheng Li

Abstract

Gaussian process models typically contain finite dimensional parameters in the covariance function that need to be estimated from the data. We study the Bayesian fixed-domain asymptotics for the covariance parameters in spatial Gaussian process regression models with an isotropic Matern covariance function, which has many applications in spatial statistics. For the model without nugget, we show that when the dimension of the domain is less than or equal to three, the microergodic parameter and the range parameter are asymptotically independent in the posterior.

While the posterior of the microergodic parameter is asymptotically close in total variation distance to a normal distribution with shrinking variance, the posterior distribution of the range parameter does not converge to any point mass distribution in general. For the model with nugget, we derive new evidence lower bound and consistent higher-order quadratic variation estimators, which lead to explicit posterior contraction rates for both the microergodic parameter and the nugget parameter. We further study the asymptotic efficiency and convergence rates of Bayesian kriging prediction. All the new theoretical results are verified in numerical experiments and real data analysis.

IP16: "Asymptotic Theory in High-Dimensional Spatiotemporal Data Analysis"

08:30 - 10:10 Sunday, 7th January, 2024
OA239 - Room 239, Old Arts Building
Chairs Yan Liu
Organiser: Yan Liu

Mengxi Yi: Robust Regularized Covariance Matrix Estimation

Koji Tsukuda: Asymptotic distribute

on of the trace of the products of four high-dimensional Wishart matrices and its application

Zeqin Lin: Asymptotic distribution of spiked eigenvalues in the large signal-plus-noise model

Kou Fujimori: Empirical likelihood methods for matrix-valued time series with long memory

IP16.01 Robust Regularized Covariance Matrix Estimation

Your name

Mengxi Yi

Abstract

We introduce a class of regularized M-estimators of multivariate scatter and show, analogous to the popular spatial sign covariance matrix (SSCM), that they possess high breakdown points. We also show that the SSCM can be viewed as an extreme member of this class. Unlike the SSCM, this class of estimators takes into account the shape of the contours of the data cloud when down-weighing observations. We also pro- pose a median based cross validation criterion for selecting the tuning parameter for this class or regularized M-estimators. This cross validation criterion helps assure the resulting tuned scatter estimator is a good fit to the data as well as having a high breakdown point. A motivation for this new median based criterion is that when it is optimized over all possible scatter parameters, rather than only over the tuned candidates, it results in a new high breakdown point affine equivariant multivariate scatter statistic.

IP16.02 Asymptotic distribution of the trace of the products of four high-dimensional Wishart matrices and its application

Your name

Koji Tsukuda

Abstract

Some properties of the Wishart distribution are provided. In particular, as for the trace of the products of four independent Wishart matrices, the explicit form of its variance is derived, and its asymptotic normality under a high-dimensional regime is shown. As an application, we propose a statisitcal test procedure for the common principal components hypothesis on the two population covariance matrices when high-dimensional variables are observed. Performances of the proposed test are numerically examined.

This study is a joint work with Professor Shun Matsuura (Keio University). This presentation is based on the article: Tsukuda and Matsuura (2021, Journal of Multivariate Analysis 186, 104822).

IP16.03 Asymptotic distribution of spiked eigenvalues in the large signal-plus-noise model

Your name

Zeqin Lin

Abstract

In this talk I will discuss the asymptotic joint distribution of the spiked eigenvalues associated with large signal-plus-noise models. Motivated by the increasing prevalence of its applications in high-dimensional statistics, we examine the scenario where the dimensionality and sample size are comparably large. Notably, our analysis reveals that the asymptotic distributions demonstrate nonuniversality, showing dependence on the distributions of the noise variables. This contrasts with what has previously been established for the spiked eigenvalues in the context of spiked population models. We also explore the application of these findings in detecting mean heterogeneity of data matrices.

IP16.04 Empirical likelihood methods for matrix-valued time series with long memory

Your name

Kou Fujimori

Abstract

The matrix-valued time series data are often observed in many fields such as economics. To analyze such data, some models, for example, matrix autoregressive models and matrix factor models have been considered. However, such models are considered only in short memory settings, and statistical inferences are mainly studied in the time domain. In this talk, on the other hand, estimation and testing problems in the frequency domain for matrix-valued time series with long-term dependence structure are considered. Using vectorization and considering the frequency domain of the time series, we develop the empirical likelihood method to construct the estimator for the unknown parameter and model diagnostics.

IP32: "Singular Stochastic Differential Equations"

08:30 - 10:10 Sunday, 7th January, 2024
EEBrown - Brown Room, Electrical Engineering Building
Chairs Xicheng Zhang
Organiser: Xicheng Zhang

Yingchao Xie: Asymptotic Behavior of Slow-Fast Stochastic Differential Equations

Jian Wang: Schr\"{o}dinger operators with with decaying potentials

Wei Liu: Well-posedness and Asymptotics of MVSPDEs

Xicheng Zhang: SDEs with supercritical distributional drifts

IP32.01 Asymptotic Behavior of Slow-Fast Stochastic Differential Equations

Your name

Yinchao Xie

Abstract

In this talk, we first introduce some background of slow-fast stochastic system. Secondly, we present our recent results on the averaging principle for slow-fast stochastic system, more precisely, it covers a class of slow-fast SDEs and SPDEs. Finally, we give the main ideas of the proofs about the optimal convergence orders.

IP32.02 Schr\"{o}dinger operators with with decaying potentials

Your name

Jian Wang

Abstract

We establish two-sided heat kernel estimates for full time and space of the Schr\"odinger operator $-\frac{1}{2}\Delta+V$ on $\R^d$, where the potential $V(x)$ is locally bounded and behaves like $c|x|^{-\alpha}$ near infinity for $\alpha\in (0,2)$ with $c> 0$, or for $\alpha>0$ with $c<0$. In particular, the potential V is decaying near infinity but does not necessarily belong to the so-called Kato-class. Our results greatly improve all known results in the literature, and it seems that the current paper is the first one in that two-sided heat kernel bounds for the long range negative potential potentials are established.

IP32.03 Well-posedness and Asymptotics of MVSPDEs

Your name

Wei Liu

Abstract

In this talk we mainly present some well-posedness and asymptotic results for a class of Mckean-Vlasov SPDEs and multi-scale stochastic systems, in particular, we show the averaging principle, large deviations principle and central limit type theorem for multiscale SPDEs and MVS(P)DEs.

IP32.04 SDEs with supercritical distributional drifts

Your name

Xicheng Zhang

Abstract

Let $d\geq 2$. In this paper, we investigate the following stochastic differential equation (SDE) in ${\mathbb R}^d$ driven by Brownian motion

$$

{\rm d} X_t=b(t,X_t){\rm d} t+\sqrt{2}{\rm d} W_t,

$$

where b belongs to the space ${\mathbb L}_T^q \mathbf{H}_p^\alpha$ with $\alpha \in [-1, 0]$ and $p,q\in[2, \infty)$, which is a distribution-valued and divergence-free vector field.

In the subcritical case $\frac dp+\frac 2q<1+\alpha$, we establish the existence and uniqueness of a weak solution to the integral equation:

$$

X_t=X_0+\lim_{n\to\infty}\int^t_0b_n(s,X_s){\rm d} s+\sqrt{2} W_t.

$$Here, $b_n:=b*\phi_n$ represents the mollifying approximation, and the limit is taken in the L^2-sense.In the supercritical case $1+\alpha\leq\frac dp+\frac 2q<2+\alpha$, if the initial distribution has an L^2-density, we show the existence of weak solutions as well as the associated Markov processes.Furthermore, if it is additionally assumed that $b=b_1+b_2+\div a$, where $b_1\in {\mathbb L}^\infty_T{\mathbf B}^{-1}_{\infty,2}$, $b_2\in {\mathbb L}^2_TL^2$ and a is a bounded antisymmetric matrix-valued function, we also establish the convergence of mollifying approximation solutions without the need to subtract a subsequence. Several examples of Gaussian random fields are provided to illustrate our results.

CP09: Contributed Paper Session

08:30 - 10:10 Sunday, 7th January, 2024
OA107 - Room 107, Old Arts Building
Chairs Yinan Lin

Xiaocong Xu: The distribution of ridgeless least squares interpolators

Yinan Lin: Logistic Regression and Classification with non-Euclidean Covariates

Moonsoo Jang: A finely tuned deep transfer learning algorithm to compare outsole images

Aoqi Zuo: Counterfactual Fairness with Partially Known Causal Graph

Arvind Kumar Nath: L\'evy Flows and associated Stochastic PDE's

CP09.01 THE DISTRIBUTION OF RIDGELESS LEAST SQUARES INTERPOLATORS

Your name

Xiaocong XU

Abstract

The Ridgeless minimum L2-norm interpolator in overparametrized linear regression has attracted considerable attention in recent years. While it seems to defy conventional wisdom that overfitting leads to poor prediction, recent research reveals that its norm minimizing property induces an `implicit regularization' that helps prediction in spite of interpolation. This renders the Ridgeless interpolator a theoretically tractable proxy that offers useful insights into the mechanisms of modern machine learning methods.

This talk takes a different perspective that aims at understanding the precise stochastic behavior of the Ridgeless interpolator as a statistical estimator. Specifically, we characterize the distribution of the Ridgeless interpolator in high dimensions, in terms of a Ridge estimator in an associated Gaussian sequence model with positive regularization, which plays the role of the prescribed implicit regularization observed previously in the context of prediction risk. Our distributional characterizations hold for general random designs and extend uniformly to positively regularized Ridge estimators.

Our theory also provides certain further conceptual reconciliation with the conventional wisdom: given any (regular) data covariance, for all but an exponentially small proportion of the signals, a certain amount of regularization in Ridge regression remains beneficial across various statistical tasks including (in-sample) prediction, estimation and inference, as long as the noise level is non-trivial. Surprisingly, optimal tuning can be achieved simultaneously for all the designated statistical tasks by a single generalized or k-fold cross-validation scheme, despite being designed specifically for tuning prediction risk.

This talk is based on joint work with Qiyang Han.

CP09.02 Logistic Regression and Classification with non-Euclidean Covariates

Your name

Yinan Lin

Abstract

We introduce a logistic regression model for data pairs consisting of a binary response and a covariate residing in a non-Euclidean metric space without vector structures. Based on the proposed model we also develop a binary classifier for non-Euclidean objects. We propose a maximum likelihood estimator for the non-Euclidean regression coefficient in the model, and provide upper bounds on the estimation error under various metric entropy conditions that quantify complexity of the underlying metric space. Matching lower bounds are derived for the important metric spaces commonly seen in statistics, establishing optimality of the proposed estimator in such spaces. A finer upper bound and a matching lower bound, and thus optimality of the proposed classifier, are established for Riemannian manifolds. We investigate the numerical performance of the proposed estimator and classifier via simulation studies, and illustrate their practical merits via an application to task-related fMRI data.

CP09.03 A finely tuned deep transfer learning algorithm to compare outsole images

Your name

Moonsoo Jang

Abstract

In forensic practice, evaluating shoeprint evidence is challenging because the differences between images of two different outsoles can be subtle. In this paper, we propose a deep transfer learning-based matching algorithm called the Shoe-MS algorithm that quantifies the similarity between two outsole images. The Shoe-MS algorithm consists of a Siamese neural network for two input images followed by a transfer learning component to extract features from outsole impression images. The added layers are finely tuned using images of shoe soles. To test the performance of the method we propose, we use a study dataset that is both realistic and challenging. The pairs of images for which we know ground truth include (1) close non-matches and (2) mock-crime scene pairs. The Shoe-MS algorithm performed well in terms of prediction accuracy and was able to determine the source of pairs of outsole images, even when comparisons were challenging. When using a score-based likelihood ratio, the algorithm made the correct decision with high probability in a test of the hypothesis that images had a common source. An important advantage of the proposed approach is that pairs of images can be compared without alignment. In initial tests, Shoe-MS exhibited better-discriminating power than existing methods.

CP09.04 Counterfactual Fairness with Partially Known Causal Graph

Your name

Aoqi Zuo

Abstract

Fair machine learning aims to avoid treating individuals or sub-populations unfavourably based on \textit{sensitive attributes}, such as gender and race. Those methods in fair machine learning that are built on causal inference ascertain discrimination and bias through causal effects. Though causality-based fair learning is attracting increasing attention, current methods assume the true causal graph is fully known. This paper proposes a general method to achieve the notion of counterfactual fairness when the true causal graph is unknown.

To select features that lead to counterfactual fairness, we derive the conditions and algorithms to identify ancestral relations between variables on a \textit{Partially Directed Acyclic Graph (PDAG)}, specifically, a class of causal DAGs that can be learned from observational data combined with domain knowledge. Interestingly, we find that counterfactual fairness can be achieved as if the true causal graph were fully known, when specific background knowledge is provided: the sensitive attributes do not have ancestors in the causal graph. Results on both simulated and real-world datasets demonstrate the effectiveness of our method.

CP46 L\'evy Flows and associated Stochastic PDE's

Your name

Arvind Kumar Nath

Abstract

We first explore certain structural properties of L\'evy flows and use this information to obtain the existence of strong solutions to a class of Stochastic PDEs in the space of tempered distributions, driven by L\'evy noise. The uniqueness of the solutions follows from Monotonicity inequality.

DL11: Distinguished Lecture Session

10:30 - 12:10 Sunday, 7th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building
Chairs Yata Kazuyoshi

Yingcun Xia: Ensemble Projection Pursuit   for General Nonparametric Regression

Effie Bura: Generalized multi-linear models for tensor-valued predictors

Qian Lin: The Optimality of Wide Neural Network in Large Dimensions

---

DL11.01 Ensemble Projection Pursuit for General Nonparametric Regression

Your name

Yingcun Xia

Abstract

The projection pursuit regression (PPR) has played an important role in the development of statistics and machine learning. However, when compared to other established methods like random forests (RF) and support vector machines (SVM), PPR has yet to showcase a similar level of accuracy as a statistical learning technique. In this paper, we revisit the estimation of PPR and propose an optimal greedy algorithm and an ensemble approach via "feature bagging", hereafter referred to as ePPR, aiming to improve the efficacy. Compared to RF, ePPR has two main advantages. Firstly, its theoretical consistency can be proved for more general regression functions as long as they are L^2 integrable, and higher consistency rates can be achieved. Secondly, ePPR does not split the samples, and thus each term of PPR is estimated using the whole data, making the minimization more efficient and guaranteeing the smoothness of the estimator. Extensive comparisons based on real data sets show that ePPR is more efficient in regression and classification than RF and other competitors. The efficacy of ePPR, as a variant of Artificial Neural Networks (ANN), demonstrates that with suitable statistical tuning, ANN can equal or even exceed RF in dealing with small to medium-sized datasets. This finding challenges the widespread belief that ANN's superiority over RF is limited to processing big data.

DL11.02 Generalized multi-linear models for tensor-valued predictors

Your name

Efstathia Bura

Abstract

We consider the problem of regressing any response on tensor-valued predictors. We assume there exists a sufficient reduction; that is, a tensor-valued function of lower dimension than the predictors' that contains the same information for the regression. We assume the distribution of the predictors given the response belongs to the quadratic exponential family in order to (a) simplify modeling and (b) keep estimation feasible. Distributions within the quadratic exponential family include the tensor normal and tensor Ising model. In this formulation, the minimal sufficient statistic of the conditional distribution of the predictors given the response is the sufficient reduction for the forward regression problem. We obtain maximum likelihood estimates of the sufficient reductions of the tensor predictors and derive the asymptotic normality of the distribution of the parameters indexing the sufficient reduction. The performance of the proposed approaches in regression and classification is compared in simulations.

390 The Optimality of Wide Neural Network in Large Dimensions

Your name

Qian Lin

Abstract

We perform a study on the generalization ability of a two-layer wide neural network for large dimensional data (where the sample size n is polynomialy depending on the dimension d of the samples, i.e., $n\asymp d^{\gamma}$ for some $\gamma >0$).

We first build a general tool to characterize the lower bound and upper bound of the kernel regression for large dimensional data through the metric entropy $\bar{\varepsilon}_{n}^{2}$ and the Mendelson complexity ε_{n}^{2} respectively.

When the target function falls into the RKHS associated with the neural tangent kernel(NTK) defined on $\bbS^{d}$, we utilize the new tool to show that the minimax rate of the excess risk of kernel regression with NTK is $n^{-1/2}$ when $n\asymp d^{\gamma}$ where $\gamma =2, 4, 8, 12, \cdots$.

We then further determine the optimal rate of the excess risk of kernel regression with NTK for all the $\gamma>0$ and find that the curve of optimal rate varying along γ exhibits several new phenomena including the multiple descent behavior and the periodic plateau behavior.

As a direct corollary, we know the above claims hold for wide neural network as well.

IP19: "Misspecification-robust Bayesians"

10:30 - 12:10 Sunday, 7th January, 2024
EEBrown - Brown Room, Electrical Engineering Building
Chairs David Frazier
Organiser: Christopher Drovandi

Renate Meyer: Bayesian nonparametric spectral analysis of locally stationary processes

Chaya Patabedi Muhamdiramalage: ABC-based Forecasting in State Space Models (joint work with Ruben Loaiza-Maya, Gael M. Martin, and David T. Frazier)

Shunan Yao: Median of Means Principle for Bayesian Inference

IP19.01 Bayesian nonparametric spectral analysis of locally stationary processes

Your name

Renate Meyer

Abstract

Based on a novel dynamic Whittle likelihood approximation for locally stationary processes, a Bayesian nonparametric approach to estimating the time-varying spectral density is proposed. This dynamic frequency-domain based likelihood approximation is able to depict the time-frequency evolution of the process by utilizing the moving periodogram previously introduced in the bootstrap literature. The posterior distribution is obtained by updating a bivariate extension of the Bernstein-Dirichlet process prior with the dynamic Whittle likelihood. Asymptotic properties such as sup-norm posterior consistency and L2-norm posterior contraction rates are presented. Additionally, this methodology enables model selection between stationarity and non-stationarity based on the Bayes factor. The finite-sample performance of the method is investigated in simulation studies and applications to real-life data sets are presented.

115 ABC-based Forecasting in State Space Models (joint work with Ruben Loaiza-Maya, Gael M. Martin, and David T. Frazier)

Your name

Chaya Weerasinghe

Abstract

Approximate Bayesian Computation (ABC) has become increasingly popular as a method for conducting both inference and forecasting in complex statistical models, most notably those in which the data generating process is intractable. In this paper we aim to explore the use of ABC to produce probabilistic forecasts in state space models (SSMs). Whilst the implementation of ABC-based forecasting to correctly-specified SSMs has been undertaken, to our knowledge there has been no explicit investigation of the empirically relevant misspecified case, and it is to that case that we give some attention. We approach the forecasting exercise by invoking recent principles of `generalized', `loss-based' or `focused' Bayesian prediction, whereby the Bayesian updating is driven by a criterion function that rewards a user-specified measure of predictive accuracy and which, in so doing, aims to produce a predictive distribution that performs well in that measure, despite possible misspecification. Two methods for incorporating the generalized criterion are investigated: one which exploits the use of an auxiliary model to produce the summary statistics that underpin the ABC method; and another in which such summary measures are eschewed, and a focused predictive is produced more directly. Our numerical results indicate that under correct specification the first approach is consistently dominant. Under misspecification, `coherent' predictions are in evidence for both methods. The dominance of one method over another is not uniform in this case, and some insights are provided as to why. An empirical application to a stochastic volatility with an intractable transition density completes the paper.

IP19.03 Median of Means Principle for Bayesian Inference

Your name

Shunan Yao

Abstract

The topic of robustness is experiencing a resurgence of interest in the statistical and machine learning communities. In particular, robust algorithms making use of the so-called median of means estimator were shown to satisfy strong performance guarantees for many problems, including estimation of the mean, covariance structure as well as linear regression. In this work, we propose an extension of the median of means principle to the Bayesian framework, leading to the notion of the robust posterior distribution. In particular, we (a) quantify robustness of this posterior to outliers, (b) show that it satisfies a version of the Bernstein-von Mises theorem that connects Bayesian credible sets to the traditional confidence intervals, and (c) demonstrate that our approach performs well in applications.

IP37: "Innovations on Time-Varying data, Network and Related "

10:30 - 12:10 Sunday, 7th January, 2024
B106 - Room 106, Babel Building
Chairs Wanjie Wang
Organiser: Wanjie Wang

Yuehaw Khoo: Randomized linear algebra for statistical problems

Yifan Cui: Instrumental variable estimation of the marginal structural Cox model for time-varying treatments

Weichi Wu: A Random Graph-based Autoregressive Model for Networked Time Series

Junhui Wang: Adaptive Merging and Efficient Estimation in Longitudinal Networks

IP37.01 Randomized linear algebra for statistical applications

Your name

Yuehaw Khoo

Abstract

In this talk, we discuss how several statistical tasks such as density estimation can be done with randomized linear algebra. This presents a method that is both computationally and statistically efficient.

IP37.02 Instrumental variable estimation of the marginal structural Cox model for time-varying treatments

Your name

Yifan Cui

Abstract

Robins (1998) introduced marginal structural models, a general class of counterfactual models for the joint effects of time-varying treatments in complex longitudinal studies subject to time-varying confounding. Robins (1998) established the identification of marginal structural model parameters under a sequential randomization assumption, which rules out unmeasured confounding of treatment assignment over time. The marginal structural Cox model is one of the most popular marginal structural models for evaluating the causal effect of time-varying treatments on a censored failure time outcome. In this paper, we establish sufficient conditions for identification of marginal structural Cox model parameters with the aid of a time-varying instrumental variable, in the case where sequential randomization fails to hold due to unmeasured confounding. Our instrumental variable identification condition rules out any interaction between an unmeasured confounder and the instrumental variable in its additive effects on the treatment process, the longitudinal generalization of the identifying condition of Wang & Tchetgen Tchetgen (2018). We describe a large class of weighted estimating equations that give rise to consistent and asymptotically normal estimators of the marginal structural Cox model, thereby extending the standard inverse probability of treatment weighted estimation of marginal structural models to the instrumental variable setting. Our approach is illustrated via extensive simulation studies and an application to estimating the effect of community antiretroviral therapy coverage on HIV incidence.

IP37.03 A Random Graph-based Autoregressive Model for Networked Time Series

Your name

Weichi Wu

Abstract

Contemporary time series data often feature objects connected by a social network that naturally induces temporal dependence involving connected neighbours. The network vector autoregressive model is useful for describing the influence of linked neighbours, while recent generalizations aim to separate influence and homophily.

Existing approaches, however, require either correct specification of a time series model or accurate estimation of a network model or both, and rely exclusively on least-squares for parameter estimation. This paper proposes a new autoregressive model incorporating a flexible form for latent variables used to depict homophily. We develop a first-order differencing method for the estimation of influence requiring only the influence part of the model to be correctly specified. When the part including homophily is correctly specified admitting a semiparametric form, we leverage and generalize the recent notion of neighbour smoothing for parameter estimation, bypassing the need to specify the generative mechanism of the network. We develop new theory to show that all the estimated parameters are consistent and asymptotically normal. The efficacy of our approach is confirmed via extensive simulations and an analysis of a social media dataset.

IP37.04 Adaptive Merging and Efficient Estimation in Longitudinal Networks

Your name

Junhui Wang

Abstract

Longitudinal network consists of a sequence of temporal edges among multiple nodes, where the temporal edges are observed in real time. It has become ubiquitous with the rise of online social platform and e-commerce, but largely under-investigated in literature. In this talk, we present an efficient estimation framework for longitudinal network, leveraging strengths of adaptive network merging, tensor decomposition and point process. It merges neighboring sparse networks so as to enlarge the number of observed edges and reduce estimation variance, whereas the estimation bias introduced by network merging is controlled by exploiting local temporal structures for adaptive network neighborhood. A projected gradient descent algorithm is proposed to facilitate estimation, where the upper bound of the estimation error in each iteration is established. Theoretical analysis of the proposed method shows that it can significantly reduce the estimation error and also provides guideline for network merging under various scenarios. We further demonstrate the advantage of the proposed method through extensive numerical experiments on synthetic datasets and a militarized interstate dispute dataset.

IP22: "Statistical Inference"

10:30 - 12:10 Sunday, 7th January, 2024
B305 - Room 305, Babel Building
Chairs Berwin Turlach
Organiser: Howard Bondell

Berwin Turlach: Extreme value copulas

Qihua Wang: A robust fusion-extraction procedure with summary statistics in the presence of biased sources

Junichi Hirukawa: Innovation algorithm of fractionally integrated (I(d)) process and applications on the estimation of parameters

IP22.01 Extreme value copulas

Your name

Berwin Turlach

Abstract

We will discuss extreme value copulas, concentrating on the bivariate case. Existing families of extreme value copulas will be reviewed and a couple of new families will be proposed.

IP22.02 A robust fusion-extraction procedure with summary statistics in the presence of biased sources

Your name

Qihua Wang

Abstract

Information from multiple data sources is increasingly available. However, some data sources may produce biased estimates due to biased sampling, data corruption, or model misspecification. This calls for robust data combination methods with biased sources. In this paper, a robust data fusion-extraction method is proposed. In contrast to existing methods, the proposed method can be applied to the important case where researchers have no knowledge of which data sources are unbiased.

The proposed estimator is easy to compute and only employs summary statistics, and hence can be applied to many different fields, e.g., meta-analysis, Mendelian randomization, and distributed systems. The proposed estimator is consistent even if many data sources are biased and is asymptotically equivalent to the oracle estimator that only uses unbiased data. Asymptotic normality of the proposed estimator is also established. In contrast to the existing meta-analysis methods, the theoretical properties are guaranteed even if the number of data sources and the dimension of the parameter diverges as the sample size increases. Furthermore, the proposed method provides a consistent selection for unbiased data sources with probability approaching one. Simulation studies demonstrate the efficiency and robustness of the proposed method empirically. The proposed method is applied to a meta-analysis data set to evaluate the surgical treatment for moderate periodontal disease and to a Mendelian randomization data set to study the risk factors of head and neck cancer.

IP22.03 Innovation algorithm of fractionally integrated (I(d)) process and applications on the estimation of parameters

Your name

Junichi Hirukawa

Abstract

The long memory phenomena frequently occur in the empirical studies of various fields. The fractionally integrated process is the one of the suitable candidate which appropriately represents the long memory property. There are two recursive algorithms for determining the one-step predictors of time series, that is, the Durbin-Levinson algorithm and the innovation algorithm. The Durbin-Levinson algorithm for the fractionally integrated process is well-known and widely used, which naturally derives the Cholesky factorization of the inverse matrix of the covariance matrix of the process. In this talk, we derive the innovation algorithm for the fractionally integrated process. The result is also applied to the derivation of the Cholesky factorization of the covariance matrix of the process in the explicit form. Moreover, the asymptotic theory of Gaussian maximum likelihood estimator (GMLE) is derived in terms of the innovation algorithm.

IP31: "Recent advances in multivariate analysis"

10:30 - 12:10 Sunday, 7th January, 2024
OA239 - Room 239, Old Arts Building
Chairs Shinpei Imori
Organiser: Shinpei Imori

Tomoyuki Nakagawa: On Robustness against outliers Bayesian estimation via γ-divergence

Takahiro Onizuka: Spatio-temporal additive model via spatial clustering and the application for the body condition analysis of common minke whales (Balaenoptera acutorostrata acutorostrata) in the Northeast Atlantic

Tomotaka Momozaki: Semiparametric Copula Estimation for Spatially Correlated Multivariate Mixed Outcomes

Shinpei Imori: On classification problem based on Fréchet distance with auxiliary variables

IP31.01 On Robustness against outliers Bayesian estimation via γ-divergence.

Your name

Tomoyuki Nakagawa

Abstract

In Bayesian statistics, the robustness against outliers is also an important issue. It is well known that an ordinary Bayesian estimator is not robust against outliers. As one of the solutions to this problem, Bayesian inference using heavy-tailed distributions has been used for a long time and has been studied extensively in recent years. However, the heavy-tailed distributions require a model to be constructed for each problem, and it is often difficult to deal with complex models. On the other hand, Bayesian estimation based on the divergence has been proposed in recent years. The divergence-based Bayesian estimation has the advantage of being flexible to various models. Numerous studies have investigated the robustness of the density power divergence-based Bayesian estimation. On the other hand, there are few studies about the robustness of the γ-divergence-based Bayesian estimation. In this presentation, we discuss the robustness of Bayesian estimation via the γ-divergence against outliers. In particular, we compare the behavior of the estimators based on the density power divergence and gamma-divergence when outliers tend to infinity. We also show the performance of the γ-divergence-based Bayesian estimation in simulation studies.

IP31.02 Spatio-temporal additive model via spatial clustering and the application for the body condition analysis of common minke whales (Balaenoptera acutorostrata acutorostrata) in the Northeast Atlantic

Your name

Takahiro Onizuka

Abstract

The additive model based on the nonparametric functions is a powerful tool to estimate unknown functions in the case assuming that the linear association in the model is unrealistic. In particular, it is critical to estimate the unknown functions adaptively if the object variable is associated by temporally and spatially non-linear effects. Therefore, in this study, we propose a spatio-temporal model based on the additive model, providing the estimates of the smoothed and interpretable spatial effect by the clustering approach and the smooth functions by the spline method. Introducing the closed form for the spline formulation, we construct an efficient algorithm. A simulation study validates the performance of the proposed method, and the model is applied to the body condition data of common minke whales (Balaenoptera Acutorostrata Acutorostrata) in the Northeast Atlantic, which is obtained by individuals caught during the period from 1993 to 2020. The data includes the blubber thickness, sex, body length, and the observed date and location. In the proposed model, we set the blubber thickness as the object variable, and estimate the effects of the other variables for the object variable nonparametrically. Through the simulation study and the real data analysis, we discuss the advantages of the proposed method.

IP31.03 Semiparametric Copula Estimation for Spatially Correlated Multivariate Mixed Outcomes

Your name

Tomotaka Momozaki

Abstract

Multivariate data having both continuous and discrete variables is known as mixed outcomes and has widely appeared in a variety of fields such as ecology, epidemiology, and climatology. In order to understand the probability structure of multivariate data, the estimation of the dependence structures among mixed outcomes is very important.

However, when location information is equipped with multivariate data, the spatial correlation should be adequately taken into account; otherwise, the estimation of dependence structures would be severely biased. To solve this issue, we propose semiparametric Bayesian inference for the dependence structure among mixed outcomes while eliminating spatial correlation. To this end, we consider a hierarchical spatial model based on the rank likelihood and a latent multivariate Gaussian process. We develop an efficient algorithm for computing the posterior using the Markov Chain Monte Carlo. We also provide a scalable implementation of the model using the nearest-neighbor Gaussian process under a large spatial dataset. We conduct a simulation study to validate our proposed procedure and demonstrate that the procedure successfully accounts for spatial correlations and correctly infers the dependence structure between variables. Furthermore, the procedure is applied to a real example collected during an international synoptic krill survey in the Scotia Sea of the Antarctic Peninsula, which includes sighting data of fin whales (*Balaenoptera physalus*), and the relevant oceanographic data.

IP31.04 On classification problem based on Fréchet distance with auxiliary variables

Your name

Shinpei Imori

Abstract

This paper considers the classification problems when the auxiliary variables are observed with the variables of interest. It is known that the classification accuracy can be improved by using the useful auxiliary variables. However, the usefulness of the auxiliary variables depends on the classification procedure. In this paper, the classifier is based on the Fréchet distance that is a distance between two Gaussian distributions. The explicit form of the Fréchet distance was derived by Dowson and Landau (1982, Journal of Multivariate Analysis), and it consists of the mean vectors and covariance matrices, which are the unknown parameters of the two Gaussian distributions. We derive the estimation accuracy of the Fréchet distance when the parameters of two Gaussian distributions are estimated from the observed data. Furthermore, we propose how to identify the useful/useless auxiliary variables for the classification problems in this situation.

IP14: "Recent Advances in Variational and Approximate Bayesian Inference"

10:30 - 12:10 Sunday, 7th January, 2024
AW153 - Room 153, Arts West North Building
Chairs Cheng Li
Organiser: Cheng Li

Pierre Alquier: Rates of convergence in Bayesian meta-learning

Ilsang Ohn: Adaptive variational Bayes: Optimality, computation and applications

Weichang Yu: Approximate Bayesian Empirical Likelihood Posterior Computation through Variational Inference

Tong T. Xin: Sampling with constraints using variational methods

---

IP14.01 Rates of convergence in Bayesian meta-learning

Your name

Pierre Alquier

Abstract

The rate of convergence of Bayesian learning algorithms is determined by two conditions: the behavior of the loss function around the optimal parameter (Bernstein condition), the probability mass given by the prior to neighborhoods of the optimal parameter.

In meta-learning, we face multiple learning tasks, that are independent but are still expected to be related in some way. For example, the optimal parameters of all the tasks can be close to each other. It is then tempting to use the past tasks to build a better prior, that we use to solve future tasks more efficiently. From a theoretical point of view, we hope to improve the prior mass condition in future tasks, and thus, the rate of convergence. In this paper, we prove that this is indeed the case. Interestingly, we also prove that we can learn the optimal prior at a fast rate of convergence, regardless of the rate of convergence within the tasks (in other words, Bernstein condition is always satisfied for learning the prior, even when it is not satisfied within tasks).

This is joint work with Charles Riou (University of Tokyo and RIKEN AIP) and Badr-Eddine Chérief-Abdellatif (CNRS). The preprint is available: https://arxiv.org/abs/2302.11709

IP14.02 Adaptive variational Bayes: Optimality, computation and applications

Your name

Ilsang Ohn

Abstract

In this paper, we explore adaptive inference based on variational Bayes. Although several studies have been conducted to analyze the contraction properties of variational posteriors, there is still a lack of a general and computationally tractable variational Bayes method that performs adaptive inference. To fill this gap, we propose a novel adaptive variational Bayes framework, which can operate on a collection of models. The proposed framework first computes a variational posterior over each individual model separately and then combines them with certain weights to produce a variational posterior over the entire model. It turns out that this combined variational posterior is the closest member to the posterior over the entire model in a predefined family of approximating distributions. We show that the adaptive variational Bayes attains optimal contraction rates adaptively under very general conditions. In addition, we characterize an implicit regularization effect of variational Bayes and show that the adaptive variational posterior can utilize this.

IP14.03 Approximate Bayesian Empirical Likelihood Posterior Computation through Variational Inference

Your name

Weichang Yu

Abstract

Bayesian empirical likelihood estimation is useful for inferring population parameters without requiring strong distributional assumptions. In this talk, we consider the computational challenges in Bayesian empirical likelihood inference that inhibits its widespread usage including a routinely observed bounded and non-convex support. Specifically, we propose a novel variational inference approach to Bayesian empirical likelihood posterior computations. This proposed approach overcomes computational challenges posed by the elaborate posterior support structure through an accurate and smooth approximation of the original empirical likelihood function. Moreover, we study an efficient stochastic gradient optimisation algorithm that utilises information from the log posterior gradient for efficient computation of the variational parameters.

Through a numerical study, the variational inference approach was confirmed to yield superior accuracy-computation time tradeoff relative to other methods. We discuss the theoretical properties of the proposed approximate posteriors including consistency and a variational Bernstein-von-Mises theorem.

IP14.04 Sampling with constraints using variational methods

Your name

Xin Tong

Abstract

Sampling-based inference and learning techniques, especially Bayesian inference, provide an essential approach to handling uncertainty in machine learning (ML). As these techniques are increasingly used in daily life, it becomes essential to safeguard the ML systems with various trustworthy-related constraints, such as fairness, safety, interpretability. We propose a family of constrained sampling algorithms which generalize Langevin Dynamics (LD) and Stein Variational Gradient Descent (SVGD) to incorporate a moment constraint or a level set specified by a general nonlinear function. By exploiting the gradient flow structure of LD and SVGD, we derive algorithms for handling constraints, including a primal-dual gradient approach and the constraint controlled gradient descent approach. We investigate the continuous-time mean-field limit of these algorithms and show that they have $O(1/t)$ convergence under mild conditions.

CP17: Contributed Paper Session

10:30 - 12:10 Sunday, 7th January, 2024
OA107 - Room 107, Old Arts Building
Chairs Changxiong Chi

Lyuyuan Zhang: Periodogram regression, a two stage mixed effects approach for modelling multiple integer-valued time series of tropical cyclone frequency

Xindong Zhao: The Best ARMA Model Group Selection and Combined Forecasting Based on Kullback-Leibler Information

Thanawan Prahadchai: Regional Frequency Analysis Based on Non-stationarity of Extreme Rainfall in South Korea

Changxiong Chi: Hill's estimator for the tail index of an ARFIMA model under heavy-tailed heteroscedastic noises

Ryota Yabe: Dickey–Fuller type test for moving average unit root

CP17.01 Periodogram regression, a two stage mixed effects approach for modelling multiple integer-valued time series of tropical cyclone frequency

Your name

Lyuyuan Zhang

Abstract

Tropical cyclones (TC) are significant indicators of evolving climate dynamics. Two primary responses of interest are the cyclone frequency and intensity. In this paper we propose a novel integrated modelling framework for simultaneous modelling of TC frequency across several meteorological regions within Australasia. The key methodological insight is to model the second order properties of multiple integer valued time series in frequency domain, instead of parametric time domain models. We take a two-stage semi parametric approach where large scale environmental variation is modelled using generalized linear models while the stochastic variation --- including spatial heterogeneity --- is estimated using spectral analysis of time series, under a hierarchical generating model. Using longitudinal data analysis we are able to jointly model periodicities in TC frequencies and their correlation with El Nino--Southern Oscillation (ENSO) cycles but also the spatial variation between regions. We project the fitted model to obtain one-step-ahead forecasts under the principles of best linear unbiased estimators.

This semi-parametric approach allows us to avoid uniqueness issues of parametric integer-valued time series modelling. Additional methodological advantages include tests for spatial heterogeneity and temporal second order stationarity. The data analysis corroborates previous findings on declining trend of tropical cyclone frequencies, in the short-term.

CP17.02 The Best ARMA Model Group Selection and Combined Forecasting Based on Kullback-Leibler Information

Your name

Xindong ZHAO

Abstract

ARMA models are widely used in the field of management science. Combined forecasting can impove the effect of forecasting. However, how to select the best model group is very important but not well done. In this paper we proposed a best model group selection method based on the Kullback-Leibler（K-L） information. First we measure the so called K-L distances between every candidate model and the true model using the K-L information, and then derive the confidence intervals of the gap between the K-L distence of each candidate model and the best model using the central limitation theory. Furthermore based on the confidence intervals we identify a group of models, which are not different significantly with the best model, as the best model group. Finally we compare the forecast ability of the best model group and the best model. The results show that the proposed method can improve the forecast with high probability when the best model is not the true model.

CP17.03 Regional Frequency Analysis Based on Non-stationarity of Extreme Rainfall in South Korea

Your name

Prahadchai Thanawan

Abstract

Climate change caused by industries is leading to non-stationarity in hydrological data. In this study, we analyzed the non-stationarity in the regional frequency analysis (RFA) of annual maximum rainfall data in South Korea. We fitted a non-stationarity model at each site and selected the best model based on the Akaike information criterion (AIC). We then clustered sites with the same best model and applied a test of homogeneity. To consider non-stationarity in RFA, we compared four methods: the conventional stationary index flood (S-IF) method, non-stationary index flood (NS-IF), non-stationary at-site frequency analysis (NS-AFA), and recommended non-stationary population index flood (NS-PIF) methods for calculating the 0.98 and 0.99 quantile, which correspond to return period 50 and 100 years for time and . We performed Monte Carlo simulations for the final regions to compare the efficiency of different quantile estimation methods. The results showed that the proposed NS-PIF method is a robust model and can solve the underestimation problem in existing S-IF and NS-IF method. In addition, it can also solve the problem of overestimation from the method. NS-AFA can also be used. The predicted return level values tend to increase at each site, which is associated with growth curves that could describe an increasing long-term predictive pattern. The findings of this study have significant implications for water management strategies and flood mitigation structure design in the country.

CP17.04 Hill's estimator for the tail index of an ARFIMA model under heavy-tailed heteroscedastic noises

Your name

Changxiong Chi

Abstract

Hill's estimator for the tail index of an autoregressive fractional integrated moving average (ARFIMA) model with unspecified and heavy-tailed heteroscedastic noises is investigated in this paper.

In particular, a self-weighted conditional-sum-of-squares estimate (SWCSSE) is first applied to estimate the parameters in the ARFIMA model. Then, based on the resulting residuals, Hill's procedure is implemented to estimate the tail index of the process. It is shown that both the SWCSSE and Hill's estimator of the tail index are consistent and asymptotically normal for all possible fractional integrated order d > −1/2. Simulations and a real-world example are also presented to illustrate the performance of the proposed method.

CP17.05 Dickey–Fuller type test for moving average unit root

Your name

Ryota Yabe

Abstract

In this talk, we will propose a Dickey-Fuller type test for unit root moving average (MA) model. The MA model with unit root is a classical time series model and has been studied extensively over the years. Furthermore, the unit root test of this model is also important in applications, for example, in testing stationarity of AR processes (Kwiatkowski et al. (1992), Saikkonen and Luukkonen (1993) and testing cointegration (Shin (1994)).

The score test (Tanaka (1990)) is a representative test for the unit root test in the MA model, but its asymptotic power deviates from that of the envelop power function and there is room for improving the power. The Wald test cannot be used because of the difficulty in deriving the asymptotic distribution of the MLE of the MA coefficient.

We propose a new test based on the test developed by Dolado, Gonzalo and Mayoral (2002) for the long memory process. Since this test does not use the asymptotic distribution of the parameter estimators to be tested, this is expected to be applicable to models where it is difficult to use the asymptotic distribution of the estimator.

The asymptotic and finite power of the proposed test has been compared with the score test, and it has been found that the both powers are almost equivalent when the alternative hypothesis is close to the null hypothesis. However, the proposed test significantly outperforms when the alternative hypothesis is more than a certain distance from the null hypothesis.

DL02: Distinguished Lecture Session

13:30 - 15:10 Sunday, 7th January, 2024
Public Lecture Theatre (PLT) - Room 122, Old Arts Building

Hsien-Kuei Hwang: Partition statistics: 10 elementary asymptotic expansions for Stirling numbers of the second kind (with a historical account)

Michael Fuchs: Gene-tree statistics: moments and limit laws for ancestral configurations

Emma Yu Jin: Permutation statistics: interactions with the theory of symmetric functions and hypergeometric series

---

DL02.01 Partition statistics: 10 elementary asymptotic expansions for Stirling numbers of the second kind (with a historical account)

Your name

Hsien-Kuei Hwang

Abstract

Stirling numbers of the second kind (or Stirling partition numbers, enumerating the number of blocks in set partitions) have been widely encountered in many different branches of mathematics and engineering sciences since the early 18th century. In probability and statistics, their uses include the classical occupancy problem, Poisson moments, minimum variance unbiased estimators, etc. Asymptotics of these numbers were first examined by Laplace in the 1780's, but his results have remained little known in the literature. In this talk, I will first review the historical developments of the asymptotics of Stirling partition numbers, and then show how a few simple ideas lead to 10 different asymptotic expansions, all proved by elementary means and covering particularly the central range (where the mode of the distribution lies). While there exist some expansions derived in the literature by elementary means (not relying on complex analysis), their proofs are incomplete. Our expansions represent the first asymptotic expansions rigorously justified by elementary approaches. I will also present another simple approach based on the sieve formula to establish the local limit theorem for these numbers. This talk is based on joint work with Chong-Yi Li and Vytas Zacharovas.

DL02.02 Gene-tree statistics: moments and limit laws for ancestral configurations

Your name

Michael Fuchs

Abstract

For a given species tree, ancestral configurations are the gene lineages which can be present at a node of the species tree over the set of all gene trees. Root ancestral configurations are the ancestral configurations which are present at the root. They play an important role in a (standard) algorithm for computing probabilities of gene trees that have evolved within the given species tree. In particular, their number is a measure for the complexity of this algorithm. In this talk, we give background, review recent results about the number of ancestral configurations and root ancestral configurations for matching species and gene trees under the uniform and Yule Harding model, and explain our recent refinements of these results which concern moments and limit laws. The talk is based on joint work with F. Disanto (University of Pisa), A. R. Paningbatan (University of the Philippines Diliman), C.-Y. Huang (National Chiao Tung University), and N. A. Rosenberg (Stanford University).

DL02.03 Permutation statistics: interactions with the theory of symmetric functions and hypergeometric series

Your name

Emma Yu Jin

Abstract

First, I will present a series of classical results on symmetric generating functions and permutation statistics, including the classical results by Carlitz, Roselle and Scoville (1966), by Garsia and Gessel (1978), by Gessel and Reutenauer (1993), etc. Second, I will focus on my recent work on the refined enumerations of permutations with respect to the Eulerian and Stirling statistics, establishing their connections to transformation formulas of basic hypergeometric series.

IP61: "Species Distribution Modeling and Capture-Recapture Models"

13:30 - 15:10 Sunday, 7th January, 2024
B106 - Room 106, Babel Building
Chairs Wen-Han Hwang
Organiser: Wen-Han Hwang

Jakub Stoklosa: Population size estimation using generalized linear latent variable models

Yang Liu: Penalized empirical likelihood estimation and EM algorithms for closed-population capture-recapture models

Yan Wang: On the Conway-Maxwell-Poisson point process

IP61.01 Population size estimation using generalized linear latent variable models

Your name

Jakub Stoklosa

Abstract

In this study, we consider a generalized linear latent variable model (GLLVM) for estimating the size of a closed population from capture-recapture data. The model allows for heterogeneity between individuals, time effects, and a latent variable component that incorporates any residual correlation across capture occasions that are not accounted for by the observed covariates (such as environmental variables) or fixed time effects. For fast estimation, we consider a reduced-rank Laplace approximation method by extending the standard Bernoulli model to a truncated Bernoulli version allowing for integration over the latent variables to estimate the regression parameters in the capture-recapture framework. We then use a Horvitz--Thompson estimator to estimate the population size. Simulations and real-data analysis are presented to demonstrate our methods.

IP61.02 Penalized empirical likelihood estimation and EM algorithms for closed-population capture-recapture models

Your name

Yang Liu

Abstract

Capture-recapture experiments are widely used to estimate the abundance of a finite population. In the presence of heterogeneity due to individual covariates, classical likelihood-based estimators may be unstable, especially when the capture probability is low. In this talk, we introduce a penalized empirical likelihood (PEL) estimation method that penalizes large abundance values. We then investigate the asymptotics of the maximum PEL estimator and the PEL ratio statistic. Additionally, we develop standard expectation-maximization (EM) algorithms for PEL to improve its practical performance. Our simulation and a real-world data analysis demonstrate that the PEL method effectively overcomes the instability of existing estimation methods and the proposed EM algorithm produces more reliable results than existing optimization algorithms.

IP61.03 On the Conway-Maxwell-Poisson point process

Your name

Yan Wang

Abstract

The Poisson point process plays a pivotal role in modeling spatial point patterns. One of its key features is that the variance and the mean of the total number of points in a given region are equal, making it unsuitable for modeling point patterns that exhibit significantly different mean and variance. To tackle such point patterns, we introduce the class of Conway-Maxwell-Poisson point processes. Our model can easily be fitted with a logistic regression, its point counts in different regions are correlated and its log-likelihood in any subregion can be easily extracted. Both simulations and real data analyses have been carried out to demonstrate the performance of the proposed model.

IP66: "Recent advances in statistical theory and applications"

13:30 - 15:10 Sunday, 7th January, 2024
B305 - Room 305, Babel Building
Chairs Kazuyoshi Yata
Organiser: Kazuyoshi Yata

Kento Egashira: Asymptotic properties of kernel k-means under high dimensional settings

Shogo Nakakita: A Langevin-type Monte Carlo method for non-log-concave non-smooth distributions

Kazuyoshi Yata: Inference on high-dimensional mean vectors by the data transformation technique

Kouji Tahata: Ordinal quasi-symmetry and its properties for multi-way contingency tables

IP66.01 Asymptotic properties of kernel k-means under high dimensional settings

Your name

Kento Egashira

Abstract

While kernel k-means has been recognized as a valuable methodology, it remains insufficiently explored from a theoretical perspective, especially in high-dimensional settings. In this talk, we aim to proceed with the current comprehension of kernel k-means. Firstly, we establish the asymptotic properties of k-means under mild and practical settings even in the context of high dimensional data. Then, we examine asymptotic properties of kernel k-means without the need to specify a particular kernel function. This allows us to investigate the difference between kernel k-means and standard k-means. Using these foundational findings, we conduct a detailed analysis of kernel k-means using the Gaussian kernel. Finally, numerical simulation studies are given and we discuss the performance of kernel k-means for high dimensional data.

IP66.02 A Langevin-type Monte Carlo method for non-log-concave non-smooth distributions

Your name

Shogo Nakakita

Abstract

We consider the approximate sampling problem from a distribution without log-concavity or smoothness. Our approach is to mollify the potential of a target distribution and sample approximately from another distribution corresponding to the mollified potential by a Langevin-type Monte Carlo method. We show the sampling complexity of the proposed method in 2-Wassertein distance.

IP66.03 Inference on high-dimensional mean vectors by the data transformation technique

Your name

Kazuyoshi Yata

Abstract

In this talk, we discuss inference problems on high-dimensional mean vectors under the strongly spiked eigenvalue (SSE) model. We precisely study the influence of the spiked eigenstructure on the inference problems using several examples. In order to remove the spiked noise, we propose a data transformation technique that avoids strongly spiked-noise spaces by precisely analyzing the huge noise structure. Using this technique, the data is transformed into the non-strongly spiked eigenvalue (NSSE) model, which enables highly accurate the inference problems on high-dimensional mean vectors. (Joint work with Prof. Makoto Aoshima (University of Tsukuba).

IP66.04 Ordinal quasi-symmetry and its properties for multi-way contingency tables

Your name

Kouji Tahata

Abstract

Square contingency tables are cross-classifications with the same row and column classifications. They are commonly used in social sciences, medical, and public health disciplines. In square contingency tables, symmetry is often more relevant than independence, as many observations concentrate on the main diagonal cells.

The ordinal quasi-symmetry model was proposed in which the log odds parameters have a linear pattern for square contingency tables with ordered categories. It is a particular case of a class of ordinal models based on f-divergence. Additionally, it is known that the symmetry model holds if and only if both the ordinal quasi-symmetry model and the mean equality model hold. The properties of goodness-of-fit test statistics between these models are obtained.

In this presentation, we discuss the ordinal quasi-symmetry model based on f-divergence for multi-way contingency tables. An information-theoretic approach shows that it is closest to symmetry under certain conditions. We introduce the model of marginal means equality for the score, which has weaker restrictions than the marginal homogeneity model. We also discuss the necessary and sufficient conditions for the symmetry model.

IP49: "BFF: Statistical foundations in the era of Data Science"

13:30 - 15:10 Sunday, 7th January, 2024
EEBrown - Brown Room, Electrical Engineering Building
Chairs Yifan Cui
Organiser: Jan Hannig

Veronika Rockova: Adaptive Bayesian Prediction Inference

Paul Edlefsen: Frequentist, Bayesian, and Dempster-Shafer Poisson regression applied to estimating HIV-1 infection times in clinical trials

Jan Hannig: A Geometric Perspective on Bayesian and Generalized Fiducial Inference

Min-ge Xie: Repro samples method for inference on discrete or non-numerical parameters

---

IP49.02 Frequentist, Bayesian, and Dempster-Shafer Poisson regression applied to estimating HIV-1 infection times in clinical trials

Your name

Paul T. Edlefsen

Abstract

We developed a Bayesian analogue of a frequentist Poisson regression approach called Poisson Fitter, and we applied this method to estimating HIV-1 infection times in two clinical trials (HVTN 704/HPTN 083 and HVTN 703/HPTN 081 Antibody Mediated Prevention [AMP] Phase 2b Prevention Efficacy Trials), using pathogen sequence data obtained during acute infection. Here we present these methods as well as analogous Dempster-Shafer (DS) methods, and discuss limitations and alternative models, including the DS Poisson Fitter and DS Binomial Fitter, and a goodness of fit procedure inspired by DS analysis. We show some applications of these methods to the problem of estimating infection times in the AMP studies, and introduce the challenge and importance of estimating HIV-1 infection times in future efficacy trials for HIV-1 vaccines and prophylactic interventions, as well as features and drawbacks of employing Bayesian and fiducial methodologies.

IP49.03 A Geometric Perspective on Bayesian and Generalized Fiducial Inference

Your name

Jan Hannig

Abstract

Post-data statistical inference concerns making probability statements about model parameters conditional on observed data. When a priori knowledge about parameters is available, post-data inference can be conveniently made from Bayesian posteriors. In the absence of prior information, we may still rely on objective Bayes or generalized fiducial inference (GFI). Inspired by approximate Bayesian computation, we propose a novel characterization of post-data inference with the aid of differential geometry. Under suitable smoothness conditions, we establish that Bayesian posteriors and generalized fiducial distributions (GFDs) can be respectively characterized by absolutely continuous distributions supported on the same differentiable manifold: The manifold is uniquely determined by the observed data and the data generating equation of the fitted model. Our geometric analysis not only sheds light on the connection and distinction between Bayesian inference and GFI, but also allows us to sample from posteriors and GFDs using manifold Markov chain Monte Carlo algorithms.

IP49.04 Repro samples method for inference on discrete or non-numerical parameters

Your name

Minge Xie

Abstract

Rapid data science developments require us to have innovative frameworks to tackle frequently seen, but highly non-trivial ``irregular inference problems,'' e.g., those involving discrete or non-numerical parameters and those involving non-numerical data, etc. This talk presents an effective and wide-reaching framework, called repro samples method, to conduct statistical inference for the irregular problems and more. We develop both theories to support our development and provide effective computing algorithms for problems in which explicit solutions are not available. The method is likelihood-free and is particularly effective for irregular inference problems.

For commonly encountered irregular inference problems that involve discrete or nonnumerical parameters, we propose a three-step procedure to make inferences for all parameters and develop a unique matching scheme that turns the disadvantage of lacking theoretical tools to handle discrete/nonnumerical parameters into an advantage of improving computational efficiency. The effectiveness of the proposed method is illustrated through case studies by solving two highly nontrivial problems in statistics: a) how to quantify the uncertainty in the estimation of the unknown number of components and make inference for the associated parameters in a Gaussian mixture; b) how to quantify the uncertainty in model estimation and construct confidence sets for the unknown true model, the regression coefficients, or both true model and coefficients jointly in high dimensional regression models. The method also has extensions to complex machine learning models, e.g., (ensemble) tree models, neural networks, graphical models, etc. It provides a new toolset for addressing the black box issues in machine learning models.

IP67: "Statistical inference"

13:30 - 15:10 Sunday, 7th January, 2024
OA239 - Room 239, Old Arts Building
Chairs John Ormerod
Organiser: Howard Bondell

Sarat Moka: Best subsets selection in regression models

John Ormerod: Moment propagation for the Bayesian lasso

Annabel Webb: A maximum penalised likelihood approach for Cox models with time-varying covariates and partly-interval censored survival data

Hongyuan Cao: A powerful empirical Bayes approach for high dimensional replicability analysis

---

IP67.01 Best Subset Selection in Linear Models via Continuous Optimization

Your name

Sarat Babu Moka

Abstract

The problem of best subset selection in linear models, such as linear regression, partial least squares, and principal component analysis, is considered with the aim of finding a fixed-size subset of features that best fits the response. This becomes particularly challenging when the total available number of features is much larger than the number of data samples. Indeed, this problem is known to be NP-hard. Existing optimal methods for solving this problem tend to be slow, while fast methods often sacrifice accuracy. In this work, I will present a novel continuous optimization method that identifies a subset solution path—a small set of models of varying size that consists of candidates for the single best subset of features, optimal in a specific sense for linear models. Our method is remarkably fast, enabling best subset selection even when the number of features exceeds thousands.

IP67.02 Moment propagation for the Bayesian Lasso

Your name

John Ormerod

Abstract

In this talk I will introduce novel approximate Bayesian Inference methodology called moment propagation. We will initially consider the case where we partition the parameter space into two components. We will then introduce a novel dimension reduction technique to extend our approach to multiple components and a simple correction to any Gaussian based posterior approximation. We apply this methodology to the Bayesian lasso where it achieves near perfect fits of the marginal posterior distributions in seconds for problems with hundreds of predictors.

IP67.04 A powerful empirical Bayes approach for high dimensional replicability analysis

Your name

Hongyuan Cao

Abstract

Identifying replicable signals across different studies provides stronger scientific evidence and more powerful inference. Existing literature on high dimensional replicability analysis either imposes strong modeling assumptions or has low power. We develop a powerful and robust empirical Bayes approach for high dimensional replicability analysis. Our method effectively borrows information from different features and studies while accounting for heterogeneity. We show that the proposed method has better power than competing methods while controlling the false discovery rate, both empirically and theoretically. Analyzing data sets from genome-wide association studies reveals new biological insights that otherwise cannot be obtained using existing methods.

IP51: "Advances in nonparametric inference for complex data"

13:30 - 15:10 Sunday, 7th January, 2024
AW153 - Room 153, Arts West North Building
Chairs Aurore Delaigle
Organiser: Aurore Delaigle

Hsin-wen Chang: Concurrent functional linear regression via plug-in empirical likelihood

Lily Wang: Distributed Heterogeneity Learning: From Spatial to Complex Data Analysis

Jinyuan Chang: Statistical inferences for complex dependence of multimodal imaging data

Germain Van Bever: Additive regression with general imperfect variables

IP51.01 Concurrent functional linear regression via plug-in empirical likelihood

Your name

Hsin-wen Chang

Abstract

Functional data with non-smooth features (e.g., discontinuities in the functional mean and/or covariance) and monotonicity arise frequently in practice. This paper develops simultaneous inference for concurrent functional linear regression in this setting. We construct a simultaneous confidence band for a functional covariate effect of interest. Along with a Wald-type formulation, our approach is based on a powerful nonparametric likelihood ratio method. Our procedures are flexible enough to allow discontinuities in the coefficient functions and the covariance structure, while accounting for discretization of the observed trajectories under a fixed dense design. A simulation study shows that the proposed likelihood ratio-based procedure outperforms the Wald-type procedure. We apply the proposed methods to studying the effect of age on the occupation time curve derived from wearable device data obtained in an NHANES study.

IP51.02 Distributed Heterogeneity Learning: From Spatial to Complex Data Analysis

Your name

Lily Wang

Abstract

Heterogeneity learning is a fundamental challenge spanning various scientific domains, from social, economic, and environmental studies to the broader landscape of complex data analysis. To address this challenge, spatially varying coefficient models have emerged as potent tools for tackling spatial regression heterogeneity. This presentation introduces a class of generalized partially linear spatially varying coefficient models that enable the inclusion of both constant and spatially varying covariate effects while balancing flexibility and parsimony. In addition, to address the challenge of extraordinarily large and complex datasets collected from modern technologies, we propose a novel distributed heterogeneity learning (DHL) method based on multivariate spline smoothing over a triangulation of the domain. The DHL algorithm has a simple, scalable, and communication-efficient implementation scheme that can almost achieve linear speedup. The DHL framework is theoretically supported by demonstrating the asymptotic normality of DHL linear estimators and DHL spline estimators' convergence rate equivalent to that of global spline estimators obtained from the entire dataset. We further expand the scope of DHL to a wider range of varying coefficient models, broadening its applicability to complex data analysis domains such as spatiotemporal data, functional data, and point cloud learning. The efficacy of the extended DHL is evaluated through comprehensive simulation studies and real-world applications.

IP51.03 Statistical inferences for complex dependence of multimodal imaging data

Your name

Jinyuan Chang

Abstract

Statistical analysis of multimodal imaging data is a challenging task, since the data involves high-dimensionality, strong spatial correlations and complex data structures. In this paper, we propose rigorous statistical testing procedures for making inferences on the complex dependence of multimodal imaging data.

Motivated by the analysis of multi-task fMRI data in the Human Connectome Project (HCP) study, we particularly address three hypothesis testing problems: (a) testing independence among imaging modalities over brain regions, (b) testing independence between brain regions within imaging modalities, and (c) testing independence between brain regions across different modalities. Considering a general form for all the three tests, we develop a global testing procedure and a multiple testing procedure controlling the false discovery rate. We study theoretical properties of the proposed tests and develop a computationally efficient distributed algorithm. The proposed methods and theory are general and relevant for many statistical problems of testing independence structure among the components of high-dimensional random vectors with arbitrary dependence structures. We also illustrate our proposed methods via extensive simulations and analysis of five task fMRI contrast maps in the HCP study.

IP51.04 Additive regression with general imperfect variables

Your name

Germain Van Bever

Abstract

An additive model is introduced, where the response variable is Hilbert-space-valued, and predictors are multivariate Euclidean, and both are possibly imperfectly observed. Considering Hilbert-space-valued responses allows us to cover Euclidean, compositional, functional and density-valued variables. By treating imperfect responses, functional variables taking values in a Riemannian manifold and the case where only a random sample can be covered from a density-valued response is available. Dealing with imperfect predictors allows us to cover various principal component and singular component scores obtained from Hilbert-space-valued variables. The smooth back-fitting method is used to estimate the additive model having such variables. Asymptotic properties of the regression estimator are provided, and a numerical study is presented.

CP06: Contributed Paper Session

13:30 - 15:10 Sunday, 7th January, 2024
BG03 - Room G03, Babel Building
Chairs Gan Yuan

Chaehyun Ryu: On a Directional Regression for Large Scale Dataset

Jeyong Lee: On the Model Selection Consistency for High-Dimensional Bayesian Poisson Regression

Sayantan Paul: Posterior Contraction Rate and Asymptotic Bayes Optimality for One Group Global-Local Shrinkage Priors in Sparse Normal Means Problem

Gan Yuan: Efficient Estimation of the Central Mean Subspace via Smoothed Gradient Outer Products

CP06.01 On a Directional Regression for Large Scale Dataset

Your name

Chaehyun Ryu

Abstract

Directional regression is an effective dimension reduction approach for capturing inherent characteristic in regression problems. We extend the idea of directional regression to handle massive datasets. In particular, this paper adopts a "divide and conquer" strategy, breaking down the dataset into manageable chunks, and subsequently merging the results based on the proximity between dimension reduction subspaces. We further harness the capabilities of capturing distance to significantly enhance computational efficiency and optimize memory usage. We demonstrate the competitiveness of our approach through a comprehensive numerical study as well as its application to a real world dataset. In both simulation and application, R packages "foreach" and "bigmemory" are utilized for optimizing the execution speed and managing memory when dealing with a massive dataset. The comparison between the proposed methodology, BIG-DR, and existing methods, namely BIG-SIR and BIG-SAVE, was conducted with a focus on computational speed and accuracy. The application of BIG-DR to real datasets demonstrates its practical applicability and versatility.

CP06.02 On the Model Selection Consistency for High-Dimensional Bayesian Poisson Regression

Your name

Jeyong Lee

Abstract

We study high-dimensional Poisson regression and the model selection consistency of its associated Bayesian model. While the Bayesian literature has extensively covered linear regression, theoretical insights into generalized linear models (GLM) have been sparse. Specifically, the model selection consistency of GLMs with sub-exponential density functions has not been widely studied, primarily because of the substantial theoretical challenges they present, especially in comparison to sub-Gaussian cases. In this paper, we delve into a Bayesian Poisson regression model with a sub-exponential density, demonstrating that this model can achieve model selection consistency.

CP06.03 Posterior Contraction Rate and Asymptotic Bayes Optimality for One Group Global-Local Shrinkage Priors in Sparse Normal Means Problem

Your name

Sayantan Paul

Abstract

We consider the well-known normal means model under sparsity, where the level of sparsity is unknown. We investigate some optimality properties of inference on the mean vector using a broad class of one-group global-local shrinkage priors including the horseshoe. Such a class of priors was considered earlier by Ghosh and Chakrabarti (2017). We show that the empirical Bayes posterior distribution of the mean vector contracts around the true mean vector as well as the empirical Bayes estimate at a near minimax rate with respect to the squared L2 loss. We prove similar results in the full Bayes approach when a non-degenerate prior is assigned on the global parameter. These results generalize those of Van Der Pas et al. (2014) and van der Pas et al. (2017) for a broad class of priors. Next under a simultaneous hypothesis testing problem, we investigate asymptotic decision theoretic optimality of a testing rule based on a full Bayes approach using our chosen class of one-group priors.

Under the asymptotic framework of Bogdan et al. (2011), we prove that the Bayes risk of our full-Bayes decision rule asymptotically attains the risk of the Bayes Oracle defined in Bogdan et al. (2011), up to some multiplicative constant. This is the first optimality result of its kind in this problem in the full Bayes approach. Our result is adaptive in nature and reinforces the argument that appropriately designed one-group shrinkage priors can be very reasonable alternatives to two-groups priors for inference in sparse problems.

CP11.03 Efficient Estimation of the Central Mean Subspace via Smoothed Gradient Outer Products

Your name

Gan Yuan

Abstract

We consider the problem of sufficient dimension reduction (SDR) for multi-index models. The estimators of the index space in prior works either have slow convergent rates or rely on stringent distributional conditions. In this paper, we show that the index space can be recovered at a fast parametric convergence rate of $C_d \cdot n^{-1/2}$ via estimating the expected smoothed gradient outer product, for a general class of distribution P_X admitting Gaussian or heavier distributions. When the link function is a polynomial with a degree at most p, and P_X is standard Gaussian, we show that C_d is proportional to d^p, which matches approximately the lower bound for learning such low-rank polynomial with any correlational statistical queries algorithms.

CP04: Contributed Paper Session

13:30 - 15:10 Sunday, 7th January, 2024
OA107 - Room 107, Old Arts Building
Chairs Saritha Kodikata

Yi Xue: Conditional maximum likelihood under two-phase sampling design

Yidi Deng: StableMate: a new statistical method to select stable predictors in omics data

Saritha Kodikata: Projection-based sequential networks for longitudinal microbiome data

John P Nolan: Sample path estimates of non-Newtonian capacity

CP04.01 Conditional maximum likelihood under two-phase sampling design

Your name

Yi Xue

Abstract

In general, sample surveys do not include all units in the target population. Ideally, we would like to make inferences about the target population. Two-phase sampling as a sample design is widely used in regression settings that involve costly covariate measurements. Through the two phases, information regarding the study of interest and auxiliary variable is collected. At the same time, our responses or observations are correlated and binary which is not handled by classic GLMs. It is clear that we need an extension of GLM to the models where suitable for correlated and categorical data. We extend this approach to generalized linear mixed models to model the sampling probabilities. The conditional maximum likelihood is an appropriate estimation method that avoids modeling the covariate distribution. The resulting likelihood function is a fraction of averages over the random effects, with integrals that do not have a closed-form in both the numerators and denominators and may involve high-dimensional integrals that cannot be evaluated analytically. When the exact likelihood function is hard to compute, approximation becomes one of the natural alternatives. A well-known method is Laplace approximation to make the estimation computationally feasible. We apply this approximation into the integrals within the likelihood to estimate those parameters of interest. We conduct an extensive simulation study to demonstrate the finite sample properties of our estimators. Different parameters settings are varied considerably to test the accuracy of proposed method and figure out the changes of influential factors related to the sample selection.

CP04.02 StableMate: a new statistical method to select stable predictors in omics data.

Your name

Yidi Deng

Abstract

Inferring reproducible relationships between biological variables remains a challenge in the statistical analysis of omics data where p > 10,000 and n < 500. Methods that identify statistical associations lack interpretability or reproducibility. We can address these limitations by inferring stable associations that are robust to external perturbation on the data. Stable associations can be an implication in causality since causal relationships are necessarily stable in some sense (Pearl et al. 2009). Unstable associations can also be of interests in certain biological applications to study functional heterogeneity in a biological system.

We developed a new regression framework, StableMate based on the concept of stabilised regression (SR), which utilise heterogenous data to enforce stability (Pfister et al. 2021). Given datasets generated from different environments, such as experiments or disease states, StableMate 1. identifies stable predictors with consistent functional dependency with the response across environments. 2. builds a robust regression model with stable predictors to enforce generalisable prediction in unseen environments. The ultimate aim is to build selection ensembles. However, unlike SR that selects stable predictors by performing stability tests on every possible predictor subset, StableMate optimizes efficiency with a greedy search based on our improved stochastic stepwise selection algorithm. In a simulation study, we show that StableMate outperformed SR for both variable selection and prediction and significantly reduces running time. In three case studies of cancer with different omics data types, we show that StableMate can also address a wide range of biological questions.

CP04.04 Projection-based sequential networks for longitudinal microbiome data

Your name

Saritha Kodikata

Abstract

The human microbiome has traditionally been studied through cross-sectional studies, offering a snapshot that limits our understanding. Recent advances in sequencing technologies allow us to explore dynamic and temporal variations through longitudinal studies. One critical analytical objective in longitudinal microbiome research is the inference of microbial association networks. Understanding these dynamic associations is crucial for uncovering the mechanistic role of microorganisms. However, conventional statistical metrics such as correlation are unsuitable due to data characteristics like compositionality, high dimensionality, and sparsity. To overcome these challenges, several statistical methods have been proposed, but for single-time-point analysis. Here, we are building upon two main ideas - conditional independence and low-dimensional data representation. Partial correlation measures the relationships between microorganisms while controlling for others. We first use Partial Least Squares (PLS) to model the current time point data based on past data to eliminate the influence of controlled microorganisms. The results from the PLS model can then be used to get a low-dimensional representation of the current data for the controlled microorganisms. These low-dimensional representations can then be used to calculate the partial correlations for the pair of microorganisms being studied. To validate our method, we applied it to simulated data, and our results demonstrated that our method outperforms popular single-time-point network methods. Furthermore, we have applied our method to three case studies and were able to extract biologically meaningful insights.

CP04.05 Sample path estimates of non-Newtonian capacity

Your name

John P Nolan

Abstract

We extend sample path methods for estimating Newtonian capacity of a d-dimensional set K in R^d to alpha-capacity, 0 < alpha < 2. The method uses paths of alpha-stable processes to obtain an empirical equilibrium measure. The alpha-capacity can be computed from this and novel methods of obtaining a confidence interval for the resulting estimate is giving.

IP55: "Topics in Statistical Inference"

15:30 - 17:10 Sunday, 7th January, 2024
EEBrown - Brown Room, Electrical Engineering Building
Chairs Karim Seghouane
Organiser: Karim Seghouane

Wei Huang: Nonparametric estimation of the continuous treatment effect with measurement error

Inge Koch: Cross-Validation for Supervised Learning with Tuning Parameter

Aurore Delaigle: Estimation of the density of a long-term trend from repeated semi-continuous data, with applications to episodically consumed food

Wenjing Yang: New Robust Canonical Correlation Approaches via alpha-divergences and Application to High-Dimensional Datasets

IP55.01 Nonparametric estimation of the continuous treatment effect with measurement error

Your name

Wei Huang

Abstract

We identify the generalised propensity score stabilised weight for a continuously valued error-contaminated treatment by a conditional moment equation. We then estimate the weights nonparametrically by maximising a local generalised empirical likelihood subject to an expanding set of conditional moment equations incorporated into the deconvolution kernels. Thereafter, we construct a deconvolution kernel estimator of the average dose-response function (ADRF). We derive the asymptotic bias and variance of our ADRF estimator and provide its asymptotic linear expansion, which helps conduct statistical inference. To select our smoothing parameters, we adopt the simulation-extrapolation method and propose a new extrapolation procedure to stabilise the computation. Monte Carlo simulations and a real data study illustrate the practical performance of our method.

IP55.02 Cross-Validation for Supervised Learning with Tuning Parameter

Your name

Inge Koch

Abstract

Recent advances in machine learning and data science have led to widespread adoption of complex predictive modelling. Increasing awareness of the ‘re- producibility crisis’ has led to calls for improved transparency and account- ability in scientific reporting. One important aspect of veridical data science is the robust estimation of prediction error. Availability of computational resources has led to cross-validation (CV) as a main tool for such estimation. We consider CV estimation in supervised learning for high-dimensional data and focus on linear regression and discriminant analysis approaches based on variable selection with direct dimension reduction as well as lasso-type sparsity criteria. We highlight how the same description of a method could in fact apply to any one of several different cross-validation implementations. We outline key principles underpinning good cross-validation practice, several ‘pitfall’ implementations which subtly violate these principles in different ways as well as a more complex and computationally intensive implementation which does not. We demonstrate the differences in the estimated error resulting from these different implementations with real data relating to endometrial cancer, in the context of high-stakes decision making where accurate and robust estimation of prediction error is critical. We use simulated data to illustrate how these different implementations result in estimators for prediction error with very different properties and relationships to the true prediction error. We call for increased detail in method-reporting, present principles for good practice in the implementation of cross-validation and make recommendations to guide cross-validation implementation.

IP55.03 Estimation of the density of a long-term trend from repeated semi-continuous data, with applications to episodically consumed food

Your name

Aurore Delaigle

Abstract

In this talk we consider semiparametric and nonparametric estimation of the density of the long-term trend of a semi-continuous variable observed repeatedly over time. Variables of this type arise when measuring the intensity of an intermittent phenomenon, such as the intake of an episodically consumed nutrient measured through repeated 24 hour recalls or the concentration of an intermittent toxic substance: on a day where the phenomenon is absent, the measurement is equal to zero; otherwise, it takes a positive value. Unlike daily consumed nutrients which are typically represented by a classical measurement error model, data with clumping at zero are usually represented by a two-part model describing the zeros and the nonzeros separately, and connected through latent variables. Several variants of the model and methods have been proposed under parametric assumptions. We study more flexible non and semiparametric approaches to this problem, which can be viewed as one with a combination of measurement errors and excess zeros.

IP55.04 New Robust Canonical Correlation Approaches via alpha-divergences and Application to High-Dimensional Datasets

Your name

Wenjing Yang

Abstract

Canonical correlation analysis (CCA) is a classical multivariate statistical method for identifying associations between two data sets. The algebraic solution is based on the sample covariance matrix, which is sensitive to outliers. In this paper, we propose two robust canonical correlation methods. The first one is motivated by the maximum likelihood estimates (MLEs) of correlated Gaussian vectors leading to CCA directions. Finding MLEs is equivalent to minimizing the Kullback-Leibler divergence between empirical distribution function and Gaussian distribution when the sample size is sufficiently large. From this point of view, we consider the generalization of the KL divergence which is a class of divergence with a single parameter α, α-divergence.

CCA solutions can be sequentially obtained by minimizing quadratic loss and adjusting associations explained by the previous estimated canonical pairs. We change the quadratic loss function by α-loss function to obtain the second robust CCA method. The consistency of estimated robust canonical directions is established under mild conditions. Also, our proposed two robust CCA methods perform better than some existed robust CCA methods in both simulated and real high-dimensional data.

IP33: "Mean field interacting particle systems and McKean-Vlasov equations"

15:30 - 17:10 Sunday, 7th January, 2024
OA239 - Room 239, Old Arts Building
Chairs Wei Liu
Organiser: Wei Liu

Zhenfu Wang: Quantitative Propagation of Chaos for 2D Viscous Vortex Model on the Whole Space

Kai Du: Empirical approximation to invariant measures of McKean--Vlasov dynamics

Longjie Xie: Long time behavior of non-linear stochastic system: the autonomous approximation method

Wei Liu: Long time behaviors of mean-field interacting particle systems and McKean-Vlasov equations

---

IP33.01 Quantitative Propagation of Chaos for 2D Viscous Vortex Model on the Whole Space

Your name

Zhenfu WANG

Abstract

We derive the quantitative estimates of propagation of chaos for the large interacting particle systems in terms of the relative entropy between the joint law of the particles and the tensorized law of the mean field PDE. We resolve this problem for the first time for the viscous vortex model that approximating 2D Navier-Stokes equation in the vorticity formulation on the whole space. We obtain as key tools the Li-Yau-type estimates and Hamilton-type heat kernel estimates for 2D Navier-Stokes on the whole space. This is a joint work with Xuanrui Feng from Peking University.

IP33.02 Empirical approximation to invariant measures of McKean--Vlasov dynamics

Your name

Kai Du

Abstract

This work reveals that the invariant probability measure of a McKean-Vlasov process can be approximated by the empirical measures of some processes including itself. These processes are described by distribution dependent or empirical measure dependent stochastic differential equations constructed from the equation for the McKean-Vlasov process. Convergence of empirical measures is characterized by upper bound estimates for their Wasserstein distance to the invariant measure. Numerical examples are given.

IP33.03 Long time behavior of non-linear stochastic system: the autonomous approximation method

Your name

Longjie Xie

Abstract

In this talk, we focus on the long time behavior of the McKean-Vlasov stochastic differential equations. We introduce a sequence of autonomous approximation systems, and with these approximation systems, we shall study:

- the existence of invariant measures (possible exhibiting phase transitions) of the original system with singular coefficients;
- uniqueness and various ergodicity (with estimates on the rate of convergence) of the invariant measure;
- phase transitions and basins of attraction of certain non-linear models.

IP33.04 Long time behaviors of mean-field interacting particle systems and McKean-Vlasov equations

Your name

Wei Liu

Abstract

In this talk, we will present our recent studies about the long time behaviors of mean-field interacting particle systems and the McKean-Vlasov equations, by using two different methods: coupling method and functional inequalities. This talk is based on the joint works with Arnaud Guillin, Liming Wu and Chaoen Zhang.

IP57: "Statistical analysis of dependent, high dimensional and massive data"

15:30 - 17:10 Sunday, 7th January, 2024
AW153 - Room 153, Arts West North Building
Chairs Ming-Yen Cheng
Organiser: Ming-Yen Chen

Xinyuan Song: Order selection for regression-based hidden Markov model

Alan Wan: A Simple Divide-and-Conquer-based Distributed Method for the Accelerated Failure Time Model

Shiqing Ling: Screening Predictors in High-Dimensional Time-Series Data

Tiejun Tong: Regularized t distribution: definition, properties and applications

---

IP57.01 Order selection for regression-based hidden Markov model

Your name

Xinyuan Song

Abstract

This study considers a regression-based hidden Markov model (RHMM) while allowing the order (i.e., the number of hidden states) to be unknown. We propose a novel likelihood-based double penalized method, along with an efficient expectation-conditional maximization with an iterative thresholding-based descent (ECM-ITD) algorithm, to perform order selection in the context of RHMM. An extended Group-Sort-Fuse procedure is proposed to rank the regression coefficients and impose penalties on the discrepancy of adjacent coefficients. The order selection consistency and convergence of the ECM-ITD algorithm are established under mild conditions. Simulation studies are conducted to evaluate the empirical performance of the proposed method. An application of the proposed methodology to a real-life study of Alzheimer's disease is presented.

IP57.02 A Simple Divide-and-Conquer-based Distributed Method for the Accelerated Failure Time Model

Your name

Alan Wan

Abstract

The accelerated failure time (AFT) model is an appealing tool in survival analysis because of its ease of interpretation, but when there is a large volume of data, fitting an AFT model and carrying out the associated inference on one computer can be computationally demanding. This poses a severe limitation for the application of the AFTmodel in the face of big data. The present paper addresses this problem by developing a simple distributed method for estimating the parameters of an AFT model based on the divide-and-conquer strategy, which has the dual benefits of statistical efficiency and computational economy. It is an iterative method that involves, for the most part,some rather simple algebraic operations, except for obtaining the initial estimate, which is based on a smoothed approximation of the Gehan estimating equation. Our results show that the proposed method yields estimates that converge after a few iterations and an estimator that is asymptotically as efficient as the benchmark estimator obtained by using the full data in one go. We also develop an associated inference procedure. The merits of the proposed method are demonstrated via an extensive simulation study.The method is applied to a kidney transplantation data set.

IP57.03 Screening Predictors in High-Dimensional Time-Series Data

Your name

Ling Shiqing

Abstract

This paper proposes a sparse vector autoregressive (AR) model for high dimensional time series in which both the dimension d and the AR order p diverge to ∞ when the sample size goes to ∞, but the model only includes s dependent variables with p_{0} lag dependence, where both s and p_0 are finite. We call the model a sp_0-sparse VAR model, and drive the convergence rate of the least squares estimators (LSE) of model parameters. Based on LSE, we construct a two-step adaptive group lasso estimation procedure and show that the procedure can screen out independent variables and determine the lag p_{0} of the remaining s dependent variables, with probability approaching 1.

The estimated parameters of the resulting VAR model, consisting of the s dependent variables, can achieve the ``oracle" properties, i.e., as efficient as the LSE when both s and p_{0} are known. The procedure is then extended to the high-dimensional factor AR model. It is showed that we can determine the number s of factors with probability approaching 1. A simulation study is carried out to assess the performance of the proposed procedure, and some real examples are given.

IP57.04 Regularized t distribution: definition, properties and applications

Your name

Tiejun TONG

Abstract

For gene expression data analysis, an important task is to identify genes that are differentially expressed between two or more groups. Nevertheless, as biological experiments are often measured with a relatively small number of samples, how to accurately estimate the variances of gene expression becomes a challenging issue. To tackle this problem, we introduce a regularized t distribution and derive its statistical properties including the probability density function and the moment generating function. The noncentral regularized t distribution is also introduced for computing the statistical power of hypothesis testing. For practical applications, we apply the regularized t distribution to establish the null distribution of the regularized statistic, and then formulate it as a regularized t-test for detecting the differentially expressed genes. Simulation studies and real data analysis show that our regularized t-test performs much better than the Bayesian t-test in the “limma” package, in particular when the sample sizes are small.

IP48: "Estimation and inference under constraints"

15:30 - 17:10 Sunday, 7th January, 2024
B106 - Room 106, Babel Building
Chairs Bradley Rava
Organiser: Eric Laber

Joseph Lawson: Bayesian Ensemble Methods for Contextual Bandit Model

Mohamed Ndaoud: Robust and Tuning-Free Sparse Linear Regression via Square-Root Slope

Bradley Rava: A Burden Shared is a Burden Halved: A Fairness-Adjusted Approach to Classification

IP48.01 Bayesian Ensemble Methods for Contextual Bandit Models

Your name

Joseph Lawson

Abstract

Contextual bandits are an important model for sequential decision making with applications ranging from clinical trials to e-commerce. While there are multiple contextual bandit algorithms which achieve optimal regret and show strong performance on benchmark problems, algorithm selection and tuning in any given application remains a major open problem. We propose the Bayesian Basket of Bandits (B3) a meta-learning algorithm which automatically ensembles a set (basket) of candidate algorithms and tuning procedures to produce a learning algorithm which dominates all algorithms in the basket. The method works by treating the evolution of a bandit algorithm as a Markov decision process in which the states are posterior distributions over model parameters and subsequently applying approximate Bayesian dynamic programming to learn an optimal ensemble. We derive both Bayesian and frequentist dominance results for the cumulative expected regret. In simulation experiments, our proposed method provides lower regret than state-of-the-art algorithms including Thompson Sampling, upper confidence bound methods, and information-directed sampling.

438 Robust and Tuning-Free Sparse Linear Regression via Square-Root Slope

Your name

Mohamed Ndaoud

Abstract

We consider the high-dimensional linear regression model and assume that a fraction of the responses are contaminated by an adversary with complete knowledge of the data and the underlying distribution. We are interested in the situation when the dense additive noise can be heavy-tailed but the predictors have sub-Gaussian distribution. We establish minimax lower bounds that depend on the fraction of the contaminated data and the tails of the additive noise. Moreover, we design a modification of the square root Slope estimator with several desirable features: (a) it is provably robust to adversarial contamination, with the performance guarantees that take the form of sub-Gaussian deviation inequalities and match the lower error bounds up to log-factors; (b) it is fully adaptive with respect to the unknown sparsity level and the variance of the noise, and (c) it is computationally tractable as a solution of a convex optimization problem. To analyze the performance of the proposed estimator, we prove several properties of matrices with sub-Gaussian rows that could be of independent interest.

IP48.03 A Burden Shared is a Burden Halved: A Fairness-Adjusted Approach to Classification

Your name

Bradley Rava

Abstract

We investigate fairness in classification, where automated decisions are made for individuals from different protected groups. In high-consequence scenarios, decision errors can disproportionately affect certain protected groups, leading to unfair outcomes. To address this issue, we propose a fairness-adjusted selective inference (FASI) framework and develop data-driven algorithms that achieve statistical parity by controlling and equalizing the false selection rate (FSR) among protected groups. Our FASI algorithm operates by converting the outputs of black-box classifiers into R-values, which are both intuitive and computationally efficient. The selection rules based on R-values, which effectively mitigate disparate impacts on protected groups, are provably valid for FSR control in finite samples. We demonstrate the numerical performance of our approach through both simulated and real data.

IP63: "Bayesian high-dimensional modeling and computation for spatial and temporal data and multivariate responses"

15:30 - 17:10 Sunday, 7th January, 2024
B305 - Room 305, Babel Building
Chairs Hsin-Hsiung Huang
Organiser: Hsin-Hsiung Huang

Hsin-Hsiung Huang: Bayesian high-dimensional variable selection for zero-inflated data

Shao-Hsuan Wang: A Bayesian framework for bilinear logistic regression variable selection

---

IP63.01 Bayesian high-dimensional variable selection for zero-inflated data

Your name

Hsin-Hsiung Huang

Abstract

Human microbiome including fungi, bacteria, viruses, and their genes has been studied for reveal interactions with infectious illness, disease states and survival status. The multivariate microbial count responses typically have excessive zeros for a specific disease or survival status. Therefore, it is challenging but important to develop a variable selection method to identify the significant biomarkers. However, there are no well-established methods for modeling multivariate zero-inflated microbiome count responses and identify significant microbiome taxa. To overcome these challenges, we propose a novel framework of Bayesian variable selection method for multivariate and high-dimensional zero-inflated count responses that identify significant taxa and reveal their associations. We apply the proposed method to both simulation and real microbiome datasets and compare it with other methods. The proposed method outperforms with respect to sensitivity, specificity, and AUC.

IP63.04 A Bayesian framework for bilinear logistic regression variable selection

Your name

Shao-Hsuan Wang

Abstract

We propose a Bayesian framework to deal with matrix-valued-covariates data via suitable shrinkage priors. This study is motivated by classifying matrix -valued data such as images. We build a bilinear logistic regression and propose an iterative Bayesian approach, which achieves strong posterior consistency. Further, we demonstrate simulation studies and applications to electroencephalography and leucorrhea data.

CP13: Contributed Paper Session

15:30 - 17:10 Sunday, 7th January, 2024
OA107 - Room 107, Old Arts Building
Chairs Ninh Tran

Ninh Tran: Adaptive Procedures for Directional False Discovery Rate Control

Yu-Hsiu Tseng: Convergence of Particle flow inspired by Wasserstein gradient flow

Mingan Yang: Bayesian semiparametric variable selection with shrinkage prior

Chuchu Wang: Nonparametric quantile scalar-on-image regression

Yudan Zou: Order selection for heterogeneous semiparametric hidden Markov models

CP13.01 Adaptive Procedures for Directional False Discovery Rate Control

Your name

Ninh Tran

Abstract

In multiple hypothesis testing, it is well known that adaptive procedures can enhance power via incorporating information about the number of true nulls present. Under independence, we establish that two adaptive false discovery rate (FDR) methods, upon augmenting sign declarations, also offer directional false discovery rate (FDR_{dir}) control in the strong sense. Such FDR_{dir} controlling properties are appealing, because adaptive procedures have the greatest potential to reap substantial gain in power when the underlying parameter configurations contain little to no true nulls, which are precisely settings where the FDR_{dir} is an arguably more meaningful error rate to be controlled than the FDR.

CP13.02 Convergence of Particle flow inspired by Wasserstein gradient flow

Your name

Yu-Hsiu Tseng

Abstract

Variational Bayesian (VB) Inference seeks to approximate a posterior distribution π by formulating the Bayesian Computation problem as an Optimisation problem. An approach is to flow a set of particles to approximates π by minimising the Kullback–Leibler (KL) divergence to π over the space of probability measures on the particle. The optimal continuity evolution of the particles' distribution is termed the gradient flow of the KL-divergence to π. The JKO scheme is a discretisation scheme to compute the gradient flow, which the results states that the limiting solution of the scheme converges to the gradient flow. At each iteration, JKO scheme computes a pushforward measure and this pushforward is determined by a vector field. In practice, such vector field is intractable and requires approximation. In this work, we describe a particle-based algorithm inspired by the JKO scheme and obtain state-of-art convergence guarantees when π is strongly log-concave for both constant step-size and diminishing step-sizes.

CP13.03 Bayesian semiparametric variable selection with shrinkage prior

Your name

Mingan Yang

Abstract

Shrinkage priors have been widely used in linear regression models for variable selection. However, they are rarely used in mixed effects models. In addition, it is commonly assumed that the random effects have normal distributions. However, substantial deviation of normal distribution might potetnially impact the ultimate variable selection models. In this article, we address the problem of joint variable selection of both fixed and random effects in nonprametric models with shrinkage priors. An efficient Gibbs sampler is developed for posterior sampling. The approach is illustrated using a simulated exampe and a real data application.

CP13.04 Nonparametric quantile scalar-on-image regression

Your name

WANG CHUCHU

Abstract

A quantile scalar-on-image regression model is developed to comprehensively study the relationship between cognitive decline and various clinical covariates and imaging factors. As a motivating example, the high-dimensional brain imaging data from the research on Alzheimer's disease are considered predictors of patients' cognitive decline. A Bayesian nonparametric model is proposed to handle the complex spatially distributed imaging data, where the coefficient function is assumed to be a latent Gaussian process. A soft-thresholding operator is introduced to capture the sparse structure of the regression coefficients. Utilizing kernel basis functions to approximate the latent Gaussian process facilitates easy-to-implement computation and consistent estimation. Inference is performed within the Bayesian framework, using an efficient Markov chain Monte Carlo algorithm. The proposed method is compared with the functional principal component analysis method in simulations and applied to a study of Alzheimer's disease.

CP13.05 Order selection for heterogeneous semiparametric hidden Markov models

Your name

Yudan Zou

Abstract

Hidden Markov models (HMMs), which can characterize dynamic heterogeneity, are valuable tools for analyzing longitudinal data. The order of HMMs, or the number of hidden states, is typically assumed to be known or predetermined by some model selection criterion in conventional analysis. As prior information about the order frequently lacks, pairwise comparisons under criterion-based methods become computationally expensive with the model space growing. A few studies have conducted order selection and parameter estimation simultaneously, but they only considered homogeneous parametric instances.

This study proposes a Bayesian double-penalized (BDP) procedure for simultaneous order selection and parameter estimation of heterogeneous semiparametric HMMs. To overcome the difficulties in updating the order, we create a brand-new Markov chain Monte Carlo algorithm coupled with an effective adjust-bound reversible jump strategy. Simulation results reveal that the proposed BDP procedure performs well in estimation and works noticeably better than the conventional criterion-based approaches. Application of the suggested method to the Alzheimer’s Disease Neuroimaging Initiative research further supports its usefulness.