Third International Conference on STATISTICS FOR TWENTY-FIRST CENTURY

Trivandrum, India
December 14-16, 2017

Invited Session

Statistical Methods for Nonlinear Data

Organizers: **Varghese George, Medical College of Georgia at Augusta University, Augusta, USA**

Ashis SenGupta, Indian Statistical Institute, Kolkata, India

Chair: **Varghese George**

1. On Efficiency and Robustness of Parameter Estimators in Langevin and Mixture Langevin Models

Somesh Kumar, Indian Institute of Technology, Kharagpur, India

Langevin distribution is a special case of rotationally symmetric distribution. We study M-estimators, restricted M-estimators and R-estimators for the location of a rotationally symmetric distribution on the unit hypersphere. The forms of the influence function and asymptotic distribution of an R-estimator are derived for a general density. Asymptotically most efficient estimators are obtained in classes of restricted M-estimators and R-estimators. In terms of gross error sensitivity, the spherical median is shown to dominate over all other estimators mentioned above under certain conditions. Explicit expressions for asymptotic relative efficiencies and gross error sensitivities of various estimators are derived for Langevin and mixture Langevin models. The trade-off between robustness and efficiency amongst various estimators has been explored.

2. Quantifying the Predictive Value of a Biomarker for Time-to-Event Outcomes

Jaya Satagopan, Memorial Sloan Kettering Cancer Center, New York, USA

Logistic regression is widely used to evaluate the association between risk factors and a binary outcome. The logistic curve is symmetric around its point of inflection. Alternative families of curves, such as additive Gompertz or Guerrero-Johnson models, have been proposed in various scenarios due to their asymmetry, as disease risk may initially increase rapidly and be followed by a longer period where the rate of growth slowly decreases. When modeling binary outcomes in relation to risk factors, an additive logistic model may not provide a good fit to the data. Suppose the outcome and an additive function of the risk factors are indeed related through an asymmetric function, but we model the relationship using a logistic function. We illustrate both from a mathematical framework and through a simulation-based evaluation that higher-order terms, such as pairwise interactions and quadratic terms, may be required in a logistic regression model to obtain a good fit to the data. Importantly, as significant higher-order terms may be a manifestation of model misspecification, these terms should be cautiously interpreted; a more pragmatic approach is to develop contrasts of disease risk coming from a good-fitting model. We illustrate these concepts in two cohort studies examining early death for late-stage colorectal and pancreatic cancer cases, and two case-control studies investigating NAT2 acetylation, smoking, and advanced colorectal adenoma and bladder cancer.

3. Identifying Non-Normally Distributed Differentially Expressed Genes

Ashis SenGupta, Indian Statistical Institute, Kolkata, India

We aim to develop an efficient test for homogeneity of mean directions of several independent circular populations, commonly termed as Analysis of Mean Directions (ANOMED), which can be universally implemented. Currently available tests have only limited applicability. In particular, tests for ANOMED are available only under highly concentrated and/or large number of groups. The present work intends to fill that gap. Focusing on the popular von Mises distribution, a simple and elegant test statistic is derived. The hurdle of the non-location-scale nuisance parameters is overcome by introducing a new approach. Second order accurate asymptotic Chi-square distribution of the test statistic is established. This test uniformly outperforms the available ones wherever they are applicable. It is a *universal* test in the sense that it continues to give satisfactory performance in the situations where the earlier tests were unusable or unsatisfactory, e.g., for highly dispersed populations. A variant of the proposed test under heterogeneous concentrations exhibited similar outperformance over LRT, the only available option in this case. Our approach is also amenable to elegant and almost straight forward generalizations to a rich variety of circular populations. This new test is illustrated through two real-life data sets.