

### Cross-Validation-Free Cramer-von Mises Nonparametric Smoothing

Bruce Brown, National University of Singapore

**Abstract:** The usual form of nonparametric smoothing involves a least-squares term and a roughness penalty, with the roughness coefficient determined retrospectively by cross-validation, based on minimising mean-squared error. However, a naive approach to testing for the correct degree of smoothing leads to a Cramer-von Mises type of criterion, which turns out to be related to both the least-squares and the roughness terms in the original formulation. This suggests a statistical way of choosing the roughness coefficient, which, fortuitously, has reliable computational properties, and which avoids the need for cross-validation.

---

### Kernel Density Estimation with Missing Data

Suzanne R. Dubnicka, Kansas State University

**Abstract:** In most parametric statistical analyses, knowledge of the distribution of the response variable, or of the errors, is important. As this distribution is not typically known with certainty, one might initially construct a histogram or estimate the density of the variable of interest to gain insight regarding the distribution and its characteristics. However, when the response variable is incomplete, a histogram will only provide a representation of the distribution of the observed data. In the AIDS Clinical Trial Study (ACT) protocol 175, for example, interest lies in CD4 counts at final follow-up, but CD4 counts collected at final follow-up are missing for more than one third of the patients. We propose methods for estimating the density of an incomplete response variable when auxiliary data are available. The proposed estimator is based on the Horvitz-Thompson estimator, and the propensity scores are estimated nonparametrically. The density estimator will be evaluated and applied to the ACT data. Extensions will also be discussed.

---

### Wild Cross-Validation for Density Estimation

Olga Y. Savchuk, Texas A&M University

Jeffrey D. Hart, Texas A&M University

Simon J. Sheather, Texas A&M University

**Abstract:** A new method of selecting the bandwidth of a kernel density estimator is proposed. The method, termed wild cross-validation, uses least squares cross-validation (LSCV) to select the bandwidth of a so-called wild kernel, and then rescales this bandwidth to be appropriate for use in a Gaussian kernel estimator. The wild kernels are linear combinations of two Gaussian kernels, and are wild since they need not be unimodal or positive. We develop theory showing that the relative error of wild CV bandwidths can converge to 0 at a rate of  $n^{-1/4}$ , which is substantially better than the  $n^{-1/10}$  rate of LSCV. The wild CV method uniformly outperforms LSCV in a simulation study and in two real data examples.

### Bruce's Adventures in Mixtureland

Bruce Lindsay, Penn State University

**Abstract:** After 29 years, I still find mixture models to be a topsy turvy playground with fun for all. I will sketch out some interesting problems I have worked on during the past couple of years. The labelling problem (with Weixin Yao) and the boundary problem (Daeyoung Kim) are what I call “straight up” mixture problems. The relationship between mixtures, modes, and clusters was described in work with Surajit Ray and Jia Li. But an offshoot (Yeojin Chung), leads to a new bias reduced kernel density estimator based on the EM algorithm. If there is time, I will jump through the looking glass and discuss a new matrix based approach to projection pursuit (Guodong Hui) which, although it was derived innocently enough, coughed up new insights when viewed from the mixture perspective.

### Penalized Clustering of Large Scale Functional Data with Multiple Covariates

Ping Ma, University of Illinois at Urbana-Champaign

**Abstract:** With the rapid advancement in high throughput technology, extensive repeated measurements have been taken to monitor the system-wide dynamics in many scientific investigations. A typical example is temporal gene expression studies, in which a series of micorarray experiments are conducted sequentially during a biological process, e.g., cell cycle microarray experiments. At each time point, mRNA expression levels of thousands of genes are measured simultaneously. Collected over time, a gene's “temporal expression profile” gives the scientist some clues on what role this gene might play during the process. A group of genes with similar profiles are often “co-regulated” or participants of a common and important biological function. Thus clustering genes into homogeneous groups is a crucial first step to decipher the underlying mechanism. In addition to the time factor, such repeated measurements often contain other covariates, e.g., replicates at each time point, species in comparative genomics studies, and treatment groups in case-control studies, as well as many factors in a factorial designed experiment. Incorporation of multiple covariates adds another layer of complexity. Clustering methods taking all these factors into account are still lacking.

In this talk, I will present a penalized clustering method for large scale data with multiple covariates through a functional data approach. Simulation studies and real-data examples are presented to investigate the empirical performance of the proposed method. Open-source code is available in the R package MFDA.

### Finding Stable Individuals for Reliability Theory: Extending 1908 Spearman-Yule Theory

Hoben Thomas, Penn State University

**Abstract:** Measure individuals twice on two occasions with the same or similar tests. The corresponding correlation coefficient  $r$  is called a reliability coefficient. Latent variables random effects models form the interpretative framework which holds that a low  $r$  indicates test deficiencies. Real differences among individuals are rarely considered as a source of difficulty.

A critical model assumption is that for all individuals, each individual's expected scores on the two occasions match. For many populations such as young children who are easily distracted, may cry, give up, or be influenced in a myriad of ways, such an assumption is patently unrealistic.

In the model proposed, each individual has their own bivariate normal distribution. If the individual's expected values (model means) match the individual is said to be stable and unstable otherwise. The problem is to identify stable individuals from the mixed population of stable and unstable individuals. Classical theory assumes all individuals are stable. The bivariate problem is transformed into a univariate difference score problem resulting in a three component univariate mixture with one stable and two unstable components. The posterior probability of an individual being stable is used in a weighted  $r$  to report a more plausible reliability. An example from infant research motivates the model.

**Data Depth and Nonparametric Multivariate Statistics:  
Spacings, Ordering and Beyond**

Regina Liu, Rutgers University  
Jun Li, University of California, Riverside  
Juan Cuesta, University of Cantabria, Spain

**Abstract:** There has been considerable new interest in nonparametric multivariate statistics due to the development of data depth and its induced center-outward ordering (or ranking) of multivariate data. We highlight some of the recent advances. In particular, we introduce and develop multivariate spacings using the order statistics derived from data depth. Specifically, the spacing between two consecutive order statistics is the region which bridges the two order statistics, in the sense that the region contains all the points whose depth values fall between the depth values of the two consecutive order statistics. These multivariate spacings can be viewed as a data-driven realization of the so-called “statistically equivalent blocks”. These spacings assume a form of center-outward layers of “shells” (“rings” in the two-dimensional case), for which the shapes of the shells follow closely the underlying probabilistic geometry. We discuss the properties and applications of these spacings. For example, we use the spacings to construct tolerance regions. The construction is nonparametric and completely data driven, and the resulting tolerance region reflects the true geometry of the underlying distribution. This is different from the existing approaches which require that the shape of the tolerance region be specified in advance. Finally, we also discuss multivariate goodness-of-fit tests based on the proposed spacings.

---

**Using Multivariate QQ-Plots to Assess Spherical Symmetry**

John I. Marden, University of Illinois at Urbana-Champaign

**Abstract:** Analogous to the univariate plots, multivariate QQ-plots can be used to compare two multivariate distributions (empirical or theoretical) by matching a set of quantiles in one distribution with the corresponding set in the other. We base our plots on Chaudhuri’s (1993) multivariate quantiles, which are vectors of the same dimension as the observations. The QQ-plots consist of arrows pointing from one distribution’s quantiles to the other’s. In two dimensions, the arrows can be plotted directly. In higher dimensions, we can look at projections. Principal component-like projections can be used to find directions in which the two distributions are most different.

These plots can be used to assess spherical symmetry of a sample by comparing the quantiles of the sample to that of a symmetrized version of the data. This technique can help in choosing the number of dimensions in principal components by stopping when the remaining variables appear spherically symmetric.

---

**Additive Noise Resistance of Location Estimators**

Ronald H. Randles, University of Florida  
Demetris Athienitis, University of Florida

**Abstract:** Estimators are compared under a model in which a symmetric signal is distorted by a small amount of additive noise. The comparisons are in terms of the local rate of change of the asymptotic bias. This robustness is determined by functional derivatives, but as a mode of comparison, it is more analogous to Pitman Relative Efficiency than it is to the Influence Function. The comparison is applied to some common affine-equivariant univariate and bivariate location estimators.

### Optimal Detection of Fechner-Asymmetry

Delphine Cassart, Université Libre de Bruxelles, Belgium  
Marc Hallin, Université Libre de Bruxelles, Belgium  
Davy Paindaveine, Université Libre de Bruxelles, Belgium

**Abstract:** We consider a general class of skewed univariate densities introduced by Fechner (1897), and derive optimal testing procedures for the null hypothesis of symmetry within that class. Locally and asymptotically optimal (in the Le Cam sense) tests are obtained, both for the case of symmetry with respect to a specified location as for the case of symmetry with respect to some unspecified location. Signed-rank based versions of these tests are also provided. The efficiency properties of the proposed procedures are investigated by a derivation of their asymptotic relative efficiencies with respect to the corresponding Gaussian parametric tests based on the traditional Pearson-Fisher coefficient of skewness. Small-sample performances under several types of asymmetry are investigated via simulations.

---

### Estimation of Hazard Rate Function under Shape Restrictions using Regression Splines

Mary C. Meyer, Colorado State University

**Abstract:** The hazard function has an important role in the understanding and modeling of survival data. It is useful to estimate the hazard function by constraining its shape; monotone or convex assumptions are often valid based on theory. Nonparametric estimation methods are proposed in which the likelihood is maximized over a set of shape-restricted regression splines. An iteratively re-weighted least squares procedure is implemented, and the estimators are shown to obtain the optimal convergence rates. Right-censoring is a simple extension.

---

### Inverse probability of censoring weighted U-statistics for right censored data with applications

Somnath Datta, University of Louisville  
Dipankar Bandyopadhyay, Medical University of South Carolina  
Glen A. Satten, Centers for Disease Control and Prevention

**Abstract:** A right-censored version of a U-statistic with a general kernel of size is introduced by the principle of a mean preserving reweighting scheme which is also applicable when the dependence between failure times and the censoring variable is explainable through observable covariates. Its asymptotic normality and an expression of its standard error are obtained through a martingale argument. Using two different kernels, we study the performance of our U-statistic by simulation and compare them with theoretical results. A doubly-robust version of this reweighted U-statistic is also introduced to preserve consistency in the face of model misspecifications. Using a Kendall's kernel, we obtain a test statistic for testing homogeneity of failure times for multiple failure causes in a multiple decrement model. The performance of the proposed test is studied through simulations. Its utility is also illustrated by applying it on a real data set.

### Use of Adaptive Robust R Procedures on Bioequivalence Type Problems

Joseph W. McKean, Western Michigan University

**Abstract:** Adaptive procedures for R estimators of regression coefficients are discussed. Several of these exploit an optimality result for R estimators. Shomrani and McKean (2003) developed an extension of an adaptive procedure for tests in simple location models as proposed by Hogg, Fisher and Randles (1974). This procedure selects one of a set of scores for the R estimation based on the residuals from an initial fit. It can be used for an analysis but, also, as a method to explore what type of scores are useful for a specified class of problems. Using this procedure and recent results for R estimators in mixed models (Kloke, McKean and Rashid, 2008), we explore the use of this R methodology for bioequivalence type problems focusing on the modeling and statistical approaches as mandated by the FDA.

---

### Consistent Model Selection and Data-driven Smooth Tests for Longitudinal Data in the Estimating Equations Approach

Lan Wang, University of Minnesota  
Annie Qu, Oregon State University

**Abstract:** Model selection for marginal regression analysis of longitudinal data is challenging due to the presence of correlation and the difficulty of specifying the full likelihood, particularly for correlated categorical data. This paper introduces a novel BIC-type model selection criterion based on the quadratic inference function (Qu, Lindsay and Li, 2000), which does not require the full likelihood or quasilielihood. With probability approaching one, the criterion selects the most parsimonious correct model. Although a working correlation matrix is assumed, there is no need to estimate the nuisance parameters in the working correlation matrix; moreover, the model selection procedure is robust against the misspecification of the working correlation matrix. The BIC-type criterion can also be used to construct a data-driven Neyman smooth test for checking the goodness-of-fit of a postulated model. This test is especially useful and often yields much higher power in situations where the classical directional test behaves poorly. The finite sample performance of the model selection and model checking procedures is demonstrated through Monte Carlo studies and analysis of a clinical trial data set.

---

### Analysis of Variance of Cross Validation Estimators of the Generalization Error of Computer Algorithms

Marianti Markatou, Columbia University  
Hong Tian, Johnson & Johnson  
George Hripcsak, Columbia University

**Abstract:** We bring together methods from statistics and machine learning to study the problem of estimating variances of cross validation estimators of the generalization error of computer algorithms. Providing an estimator of the variance of the cross validation estimator of the generalization error is a difficult problem, particularly if one wants to take into account various sources of variability. We provide a general framework which allows us to treat this problem as a problem in approximating the moments of a statistic. For the simple case of predicting the sample mean and when the loss functions are smooth, we show that the variance of the cross validation estimator of the generalization error is a function of the moments of the random variables  $Y$ ,  $Z$  where  $Y$  denotes the cardinality of the intersection of two difference training sets and  $Z$  denotes the cardinality of the intersection of the corresponding test sets. We extend these results to regression, kernel regression and classification, and illustrate the methods through a health services research example on the hospital length of stay of patients that have undergone colon cancer surgery.

### Nonparametric Approach for Firms' Default

Pasquale Cirillo, University of Bern, Switzerland

Jürg Hüsler, University of Bern, Switzerland

**Abstract:** A new intuitive approach is discussed for firms' default or generalized shock models using urn processes. This approach is not depending on a parametric model as in sum or extreme shock models or even the more generalized shock models introduced in Gut and Hüsler (2005). The urn approach allows us to indirectly model the moving risky threshold in generalized extreme shock models. This plays the important role in modeling firms' default. The basic idea is to link the types of the balls in the urn with the risk or the levels of risk a system can face. The evolution of the process is given by a triangular reinforcement matrix. Thus no parametric distribution is assumed for the risk process.

In particular, assuming that a firm can experience three different levels of risk (no risk, risk and default), we introduce a dependence among the levels, so that the probability of default increases every time the firm enters the risky state, while it decreases (but does not disappear) the more the firm spends in the non-risky one. This approach makes it possible to predict firms' default probabilities with a good degree of approximation and to obtain limit distributions that nicely reproduce the empirical results one can find in the literature. The model is applied to real data to show the prediction accuracy of the firms' default times and default probabilities and to compare it with a known benchmark prediction model.

---

### Records and Hurricanes

Sneh Gulati, Florida International University

**Abstract:** Statisticians are constantly engaged in a game with nature. Nature provides the true underlying population and the statistician tries to guess what it is based on observed data. The guessing game includes estimation and prediction, both from complete and from incomplete data. Here we present such methodology from a particular type of incomplete data called record-breaking data, that is, data generated from setting new records. We encounter records on a daily basis, e.g. sports records, meteorological records, financial records etc. Besides arising naturally in our daily lives, in many industrial quality control experiments and destructive stress testing, the only available data are successive minima (or maxima), i.e. record-breaking data. This paper will attempt to provide a comprehensive review of all the results related to the nonparametric inference from such data. A major part of the author's work has focused on smooth estimation from record-breaking data. Recently, the author has also applied the concept of smooth estimation to estimate the genesis time of hurricanes in the Atlantic Basin. The second part of the talk will focus on the use of nonparametric techniques to model some characteristics of a hurricane.

### Nonparametric Clustering on Mixtures of Functional Data

Haiyan Wang, Kansas State University

James Neil, Kansas State University

Forrest Miller, Kansas State University

**Abstract:** In this talk, I will present a method for effectively detecting unknown clusters in high-dimensional functional data. Examples of such data include gene expression levels measured over time from microarray experiments, functional magnetic resonance imaging (fMRI), mass spectrometry data from proteomics, lipidomics, etc. We define clusters through the unknown high-dimensional multivariate distributions of all observations. Kullback-Leibler information and Mahalanobis generalized squared distance can fail to provide meaningful measures of distance between distributions in such high-dimensional settings.

We propose a new similarity measure and an agglomerative clustering algorithm, called PCLUST, to effectively differentiate among high-dimensional populations. The algorithm produces invariant results under monotone transformations of data and does not require users to specify the number of clusters. Simulations show that PCLUST significantly outperforms nine other popular algorithms in both clustering accuracy and robustness. An application in identifying biomarkers using time course gene expression data from Arabidopsis in response to environmental stresses is illustrated.

### Multivariate Nonparametric Methods Based on Spatial Signs and Ranks

Hannu Oja, Tampere School of Public Health, Finland

**Abstract:** Classical multivariate statistical inference methods (Hotelling's tests, multivariate analysis of variance, multivariate regression, etc.) are based on the use of  $L_2$  criterion functions. In this talk we consider alternative  $L_1$  criterion functions with related tests and estimates. Multivariate  $L_1$  criterion functions for estimation problem are used to extend the concepts of sign and rank to the multivariate case.

We consider three different multivariate  $L_1$  criterion functions utilizing Euclidean distance. The first criterion function, the mean deviation of the multivariate residuals, is the basis for the so called least absolute deviation (LAD) methods; it yields different spatial median-type estimates and spatial sign tests in the one-sample, two-sample,  $c$ -sample and finally general linear regression settings. The second objective function is the mean difference of the residuals, and the third one is the sum of the lengths of the pairwise sums and pairwise differences of the residuals. The second and third objective functions generate multivariate Hodges-Lehmann type estimates and spatial rank and signed-rank tests for different location problems. For these ideas, see also Hettmansperger and Aubuchon (1988).

In the talk, we briefly review the theory of the multivariate spatial sign and rank methods, tests and estimates, in the one sample, several samples and, finally, multivariate linear regression cases. See Mottonen and Oja (1995) and Oja and Randles (2004). Transformation-retransformation technique is used to obtain affine invariant tests and equivariant estimates. See e.g. Hettmansperger and Randles (2003). The theory is illustrated with examples.

### Two-Sample Rank-Sum Test under Order Restricted Randomized Designs

Ömer Öztürk, Ohio State University  
Yiping Sun, Forest Research Institute, New Jersey

**Abstract:** A new nonparametric test is proposed based on order restricted randomized design (ORRD) for the location shift between two populations. The ORRD is similar to a ranked set sample and exploits the use of the subjective information on experimental units. Under the ORRD, sets of experimental units are recruited from a population along with subjective information that they may have. This subjective information is then used to create artificial covariates through judgment ranking of the experimental units. These artificial covariates with a proper randomization scheme induce a positive correlation structure among within-set response measurements. This positive correlation structure then acts as a variance reduction technique in the inference of a contrast parameter in an ORRD.

Based on the ORRD, a rank-sum test is developed. It is shown that the asymptotic null distribution of the proposed test is distribution-free even if the ranking information is not perfect. The point and interval estimates for the location shift parameter are developed. An optimal design is constructed by maximizing the asymptotic Pittman efficacy of the test. It is shown that the test outperforms its competitors even under imperfect ranking information. The use of the proposed test is illustrated in a data set from a clinical trial study.

---

### Rank-based tests of multivariate independence in independent component models

Davy Paindaveine, Université Libre de Bruxelles, Belgium  
Hannu Oja, Tampere School of Public Health, Finland  
Sara Taskinen, University of Jyväskylä, Finland

**Abstract:** The so-called independent component (IC) model states that the observed  $p$ -vector  $X$  is generated via  $X = \Lambda Z + \mu$ , where  $\mu$  is a  $p$ -vector,  $\Lambda$  is a full-rank matrix, and the centered random vector  $Z$  has independent marginals  $Z_i$ . We consider the problem of testing, on the basis of  $n$  i.i.d. copies of  $X = (X^{(1)'}, X^{(2)'})'$ , the null hypothesis under which the multivariate marginals  $X^{(1)}$  and  $X^{(2)}$  are independent. Under a symmetry assumption on the  $Z_i$ 's, we propose a class of semiparametric procedures, which are based on the componentwise signed ranks of the estimated independent components (the latter are obtained under the null via a recent procedure due to Oja et al. 2006). This componentwise signs-and-ranks methodology was first proposed by Puri and Sen (1971). However, unlike the Puri and Sen tests, our tests (i) are affine-invariant and (ii) achieve, for adequately chosen scores, local and asymptotic optimality (in the Le Cam sense) at given densities. They are also valid without any moment assumptions. Local powers and asymptotic relative efficiencies with respect to the classical Gaussian procedure (namely, Wilks' Gaussian LRT test) are derived. Finite-sample properties are investigated through a Monte Carlo study.