

# Nonparametric Statistics, Refined, Redefined, and Renewed

Abstracts – last updated April 3, 2009

---

*Arne C. Bathke*, University of Kentucky, and *Edgar Brunner*, University of Göttingen

## **Improvement and Generalization of the Box-Greenhouse-Geisser Adjustment**

Based on the observation that two seemingly unrelated methods, namely the Box-Greenhouse-Geisser adjustment for the analysis of repeated measures data (Box 1954, Geisser and Greenhouse 1959) and the ANOVA-type statistic that was developed for the nonparametric analysis of factorial designs (Brunner *et al.* 1997, 1999), are actually cousins, or perhaps even siblings, we explore the implications that this rediscovered family connection has for both. Practical consequences include, for example, the mutual suitability of dedicated software, at least in special cases. Further, we investigate how this connection could help in improving either method. This motivates new nonparametric methods for the analysis of repeated measures data. In particular, we are interested in the development of nonparametric inferential tools for situations in which the number of repeated measurements per subject is large, perhaps even larger than the number of subjects.

---

*Soutir Bandopadhyay*, Texas A&M University

## **Nonparametric Covariogram Estimation Based on Irregularly Spaced Spatial Data**

We propose a nonparametric estimator for the covariogram function based on irregularly spaced spatial data generated by a stochastic design. Unlike earlier works on the problem, the proposed estimator of the covariogram does not involve any smoothing. As a result, the proposed method avoids the vexing problem of finite sample choice of the smoothing parameter for practical applications. We investigate theoretical properties of the covariogram estimator and present results from a moderately large simulation study.

---

*Ming-Yen Cheng*, University College, London, and *Jyh-Shyang*, Tamkang University

## **Adapting to Design Sparsity in Univariate and Bivariate Local Linear Regression**

Local linear regression enjoys many nice theoretical properties such as automatic boundary correction and linear minimax optimality and has become very popular in

applications. In finite sample cases, the local least squares problem used to obtain the local linear estimator becomes ill-posed when the design is sparse and, as a result, the estimator either does not exist or exhibits drastic roughness there. We propose a new method to tackle this difficulty in both univariate and bivariate case. It is computationally simple and does not involve any extra tuning parameters, like existing methods do. We show that it has the same asymptotic mean squared error as the original local linear estimator. Numerical studies demonstrate that it has very good finite sample performance.

---

*Wei Dou*, Yale University

### **Functional Generalized Linear Models: Methodology, Convergence Rates, and Applications**

Slope function estimation in functional ordinary linear regressions has been substantially studied recently, from both theoretical and practical aspects. We consider exponential family models whose canonical parameters are specified as linear functionals of an unknown infinite dimensional slope function  $B$  in an  $L^2$  space. We established a minimax lower bound for estimation of  $B$  using  $L^2$  loss, and also construct an estimator that achieves the minimax rate. Our method adapts ideas from LeCam theory of equivalence of experiments to simplify the analysis of our estimator. The asymptotic equivalence theorem provides a powerful technique tool for general infinite dimensional estimation problems. In conclusion, we apply a functional logistic regression model and our methodology to an interesting financial data set. (Joint work with *David Pollard* and *Harrison H. Zhou*)

---

*Sam Efromovich*, University of Texas at Dallas

### **Nonparametric Regression with Missing Data**

Regression based on missing at random (MAR) responses is a familiar topic in the missing data literature, and imputation of missing responses is the remedy. What will be if inadvertently predictors, corresponding to missing responses, are also lost and only “complete-case” pairs of observations are available? It is well documented in the literature that a complete-case analysis is generally inadmissible. It is proved that, contrary to this general rule, in a nonparametric regression this approach implies the same sharp minimax mean integrated squared error (MISE) convergence as by a sharp minimax estimator based on the underlying MAR sample. Then “the table is turned around” and a less studied, even in the parametric and semiparametric literature, problem of missed at random (MAR) predictors is explored. Both lower and upper minimax bounds are presented and possible estimators are discussed.

---

*Xin Gao*, York University

### **Nonparametric Multiple Comparison Procedures for Unbalanced Experimental Designs**

In this talk, we consider nonparametric multiple comparison procedures for unbalanced one-way and two-way factorial designs under a pure nonparametric framework. For multiple comparisons of treatments versus a control concerning the main effects or the simple factor effects, the limiting distribution of the associated rank statistics is proven to satisfy the multivariate totally positive of order two condition. Hence, asymptotically the proposed Hochberg procedure strongly controls the familywise type I error rate for the simultaneous testing of the individual hypotheses. In addition, we propose to employ Shaffer's modified version of Holm's stepdown procedure to perform simultaneous tests on all pairwise comparisons regarding the main or simple factor effects. We will also discuss simultaneous tests on interaction effects. The logical constraints in the corresponding hypothesis families are utilized to sharpen the rejective thresholds and improve the power of the tests.

---

*Nader Gemayel*, Ohio State University

### **Nonparametric Estimation in Ranked Set Sampling with Imperfect Ranking Induced by a Concomitant Variable in a Finite Population Setting**

We consider the problem of nonparametric estimation in Ranked Set Sampling (RSS) with imperfect ranking induced by a concomitant variable. We will see that this estimation problem occurs in very high-dimensional parameter spaces, and that simple likelihood methods rarely zoom in on specific regions of the parameter space that agree with the data. We must therefore consider restrictions on the parameter space and different optimality criteria. In particular, we examine the interpretation, usefulness, and estimation of judgment rank distributions, and ask how we can relate two judgment order statistics of the same rank when their associated concomitant values differ significantly. We will conclude by applying our results to an appropriate data set.

---

*Marc Hallin, Davy Paindaveine, and Miroslav Šiman*, Université Libre de Bruxelles

### **Multivariate Quantiles: from $L_1$ Optimization to Halfspace Depth**

A new multivariate concept of quantile, based on a directional version of Koenker and Bassett's traditional regression quantiles, is introduced for multivariate location and multiple-output regression problems. In their empirical version, those quantiles can be computed efficiently via linear programming techniques. Consistency, Bahadur representation and asymptotic normality results are established. Most importantly, the contours generated by those quantiles are shown to coincide with the classical halfspace

depth contours associated with the name of Tukey. This relation allows not only efficient depth contour computations by means of parametric linear programming, but also for transferring from the quantile to the depth universe such asymptotic results as Bahadur representations. Finally, linear programming duality opens the way to promising developments in depth-related multivariate rank-based inference.

---

*Solomon Harrar*, University of Montana

### **Modified Rank-Based MANOVA: Asymptotics and Small Sample Approximations**

In this talk, we present results for testing main effects and interaction effects in heteroscedastic multi-factor MANOVA. In particular, we propose modifications to the usual MANOVA statistics to account for heteroscedasticity and to study their asymptotic distributions in a non-standard asymptotic setting under non-normality. The results are extended to rank-based versions of the statistics. Based on the asymptotic results, we devise some small sample approximations and evaluate their accuracies via a simulation study. A real data example is used to illustrate the application of the results.

---

*Frank Konietschke*, University of Göttingen

### **Simultaneous Confidence Intervals for Nonparametric Relative Contrast Effects**

In practice, the assumption of normality of the data is often not fulfilled. Hence there is a need for statistical procedures which can handle skewed distributions as well as ordinal or categorical data.

In the literature, there are many nonparametric analysis of variance type procedures. But the application of these procedures has disadvantages:

- (i) The global hypothesis tested by analysis of variance (quadratic tests) is, in general, not the question of practitioners.
- (ii) When the global hypothesis is rejected, then multiple comparison procedures must be applied, but the results are often not consistent with the global test.
- (iii) Most nonparametric procedures cannot provide meaningful confidence intervals for the underlying effects.

In this talk, we will present some results from my PhD thesis, trying to solve these problems. We will discuss some new so-called relative contrast effects, which can be interpreted as the underlying effects of a nonparametric multiple testing procedure. Almost surely consistent estimators and their multivariate distribution will be derived. We will present a new multiple testing approach and simultaneous confidence intervals for the effects.

---

*Denis Larocque*, HEC Montreal

### **Multivariate Nonparametric Methods for Clustered Data: A Review and Some Recent Developments**

During the last decade, there has been a wide interest to extend univariate and multivariate nonparametric procedures based on signs and ranks to clustered and hierarchical data. In this talk, I will present a review of these developments but the main focus will be on the more recent developments using the spatial signs and ranks. In particular, using the traditional mixed models notation, I will outline how multivariate nonparametric procedures for one sample and several samples problems can be extended to clustered data and present some results for general score functions. (Joint work with *Hannu Oja*, *Jaakko Nevalainen* and *Riina Haataja*, Tampere School of Public Health, University of Tampere, Tampere, Finland)

---

*Soyeon Lee*, Occidental College, *Wentao Gu*, Zhejiang Gongshang University, and *Lanh Tran*, Indiana University

### **Fixed Design Regression for General Linear Time Series**

We are concerned with recovering a regression function  $g(x)$  on the basis of noisy observations taken at design points  $x_i$ . The corresponding observations are corrupted by additive dependent noise induced by a general linear time series. The regression function is estimated by a smoother, which is shown to have an asymptotic multivariate normal distribution at multiple points. The problem of finding confidence bands for  $g(x)$  is discussed. An illustrative example is also exhibited. The results for finite samples are evaluated by computer simulations.

---

*Jiexiang Li*, College of Charleston

### **Asymptotic Normality for Deconvolution Kernel Density Estimators from Random Fields**

The talk discusses estimation of a continuous density function of a target random field  $X_{\mathbf{i}}$ ,  $\mathbf{i}$  in  $Z^N$ , which is contaminated by measurement errors. In particular, the observed random field  $Y_{\mathbf{i}}$ ,  $\mathbf{i}$  in  $Z^N$ , is such that  $Y_{\mathbf{i}} = X_{\mathbf{i}} + \varepsilon_{\mathbf{i}}$ , where the random error  $\varepsilon_{\mathbf{i}}$  is from a known distribution and independent of the target random field. Compared to existing results, improvements in two directions are obtained. First, random vectors in contrast to univariate random variables are investigated. Second, a random field with a certain spatial interaction instead of i.i.d. random variables is studied. Asymptotic normality of the proposed estimator is established under appropriate conditions.

---

*Jun Li*, University of California, Riverside, and *Regina Liu*, Rutgers University

### **Multivariate Spacings Based on Data Depth and Construction of Nonparametric Multivariate Tolerance Regions**

In this talk, we introduce and study multivariate spacings. The spacings are developed using the order statistics derived from data depth. These multivariate spacings can be viewed as a natural generalization of univariate spacings to the multivariate setting. These spacings assume a form of center-outward layers of "shells" ("rings" in the two-dimensional case), where the shapes of the shells follow closely the underlying probabilistic geometry. The properties and applications of these spacings are studied. In particular, the spacings are used to construct tolerance regions. The construction of tolerance regions is nonparametric and completely data driven, and the resulting tolerance region reflects the true geometry of the underlying distribution. This is different from most existing approaches which require that the shape of the tolerance region be specified in advance. The proposed tolerance regions are shown to meet the prescribed specifications. They are also asymptotically minimal under elliptical distributions. Finally, we present a simulation and comparison study on the proposed tolerance regions.

---

*Satyaki Mazumder*, University of Texas at Dallas

### **Asymptotic Results for Scaled-Deviation Type Outlyingness Functions, with Applications**

Using projection pursuit, interesting multivariate outlyingness functions may be constructed from the associated univariate scaled-deviation outlyingness functions taken over projections of the data onto lines. In particular, we consider univariate outlyingness of the form  $O(x) = (x - \text{median})/\text{MAD}$ , where MAD may denote the usual MAD or a variant as in Tyler (1984). The supremum of the projected outlyingness of a data point over all projections leads to the well-known "projection outlyingness". However, it is of interest to use more informative approaches that can be sensitive to the presence of several large deviations and also yield more convenient asymptotic theory. This entails using a vector of some finite number of projected scaled deviations. For such a vector, we obtain the joint asymptotic normality of the components using the classical Bahadur representation for the sample median and a new Bahadur representation for the sample MAD and its variants that removes a symmetry assumption of Hall and Welsh (1985). As in Pan, Fung, and Fang (2000), we then construct quadratic form type outlyingness functions having asymptotic chi-square distributions. Our results include two important advances: extension to allow variants of the MAD, and elimination of unnecessary symmetry and regularity assumptions. (Joint work with *Robert Serfling*)

---

*Ursula Müller, Texas A&M University*

### **Efficient Estimators for Nonlinear Regression Models with Responses Missing at Random**

The problem of incomplete data is frequently encountered in many fields, with the health sciences and survey research being obvious examples. The past two decades have seen a considerable amount of literature, but for the most part the focus has been on finite-dimensional parametric models. My current research seeks to go beyond this by allowing more flexible "semiparametric" models, and by constructing estimators that are efficient, i.e., as accurate as possible.

In this talk I will focus on regression models with responses that are allowed to be missing at random. The models are semiparametric in the following sense: I assume a parametric (linear or nonlinear) model for the regression function but no parametric form for the distributions of the variables; I only assume that the errors have mean zero and are independent of the covariates. I introduce an easy-to-implement weighted imputation estimator, adapting empirical likelihood ideas, for estimating general expectations of functions of covariate and response. The estimator is efficient in the sense of Hájek and Le Cam, since it uses all model information.

---

*Kimihiro Noguchi, University of Waterloo*

### **Combination of Levene-Type and Finite-Intersection Tests for Homogeneity of Variances Against Ordered Alternatives**

A problem of detecting monotonic increasing/decreasing trends in variances from  $k$  samples is widely met in many applications, e.g., financial data analysis, psychology, medical and environmental studies. There exists a variety of tests for homogeneity of variances against ordered alternatives. However, most of such tests rely on the assumption of normality and are not robust with respect to its violation, which eventually leads to inadequate estimation of Type I error and unreliable conclusions. Such violations are especially severe for heavy-tailed and skewed data. In this talk we propose to combine a robust Levene-type test and a finite-intersection method against ordered alternatives, which enables to relax the assumption of normality on observations and consider a more general class of data. The new combined procedure yields an accurate estimate of the size of the test and provides competitive power results. In addition, we discuss various modifications of the proposed test for a case of unbalanced design, i.e., when the number of observations is different among groups. We present theoretical justifications of our new combined test and illustrate its application by simulations and case studies. (Joint work with *Yulia R. Gel*)

---

*Omer Ozturk*, Ohio State University

### **Nonparametric Maximum Likelihood Estimation of CDF and Within-set Ranking Error Probabilities in Ranked-set Sampling**

In this talk, I will introduce estimators for the cumulative distribution function (CDF) and within-set judgment ranking error probabilities in ranked set sampling. The judgment ranking information is constructed based on models of Bohn-Wolfe (Bohn and Wolfe, 1994) and Frey (Frey, 2007). The parameters of these models and the CDF of the underlying population are estimated by maximizing a nonparametric likelihood function. A missing data model is introduced to construct an efficient computational algorithm. Asymptotic properties of the estimators are developed and efficiencies are compared with competitors in the literature. The advantages of the new estimators are that they require essentially no assumption on the underlying distribution function, that they provide an estimate of the quality of within-set ranking information, and that they lead to a valid statistical inference even under imperfect ranking. The proposed estimators are applied to a water flow data set to estimate judgment ranking information and the underlying distribution function.

---

*Davy Paindaveine*, Université Libre de Bruxelles

### **On Multivariate Runs Tests for Randomness**

This paper proposes several extensions of the concept of runs to the multivariate setup, and studies the resulting tests of multivariate randomness against serial dependence. Two types of multivariate runs are defined: (i) an elliptical extension of the spherical runs proposed by Marden (1999), and (ii) an original concept of matrix-valued runs. The resulting runs tests themselves exist in various versions, either based on the so-called spatial signs or on the hyperplane-based multivariate signs known as Randles' interdirections. All proposed multivariate runs tests are affine-invariant and highly robust: in particular, they allow for heteroskedasticity and do not require any moment assumption. Their limiting distributions are derived under the null hypothesis and under sequences of local vector ARMA alternatives. Asymptotic relative efficiencies with respect to Gaussian Portmanteau tests are computed, and show that, while Marden-type runs tests suffer severe consistency problems, tests based on matrix-valued runs perform uniformly well for moderate-to-large dimensions. A Monte-Carlo study investigates the small-sample properties of the proposed procedures. A real data example is treated and shows that combining Marden-type runs tests and tests based on matrix-valued runs may provide some insight on the reason why rejection occurs.

---

*Javier Rojo*, Rice University, and *Richard C Ott*, Mesa State College

### **Testing for Long Tails**

After a brief review of classification of distributions by tail behavior, procedures for testing the null hypothesis of medium tails versus long tails are presented. The consistency of the tests against Long- and Short-tailed alternatives is also discussed. Results from a simulation study are also presented.

---

*Robert Serfling*, University of Texas at Dallas

### **On Equivariance/Invariance Properties of Multivariate Depth and Related Functions**

In what ways should estimators and test statistics, or sample quantile, depth, and outlyingness functions, desirably transform when the data undergo transformation to another coordinate system? We will set the stage for consideration of this question by looking at the linkages, indeed the equivalence, between multivariate depth, outlyingness, quantile, and centered rank functions. It will then be clear that this question is nontrivial, because, in general, for example, quantile functions of affinely transformed distributions entail re-indexings and these must be such that simultaneously the corresponding outlyingness functions are affine invariant. A general and flexible formulation of affine equivariance/invariance for all these functions will be given. Also, connections with covariance functionals, transformation-retransformation (TR) functionals, and invariant coordinate selection (ICS) functionals, and weak versions of these functionals, will be discussed.

---

*Christopher Sroka*, Battelle Memorial Institute, Columbus

### **Approaches for Allocating Observations in a Stratified Ranked Set Sample**

Estimators of the mean based on ranked set sampling (RSS) have been shown to be more precise than the sample average from a simple random sample. Increased precision can be obtained if the population is stratified and RSS is conducted within each stratum. This method, called stratified ranked set sampling (SRSS), results in more precise estimation of the mean than stratified simple random sampling when the numbers of observations assigned to the strata are the same for both methods. This presentation describes methods for allocating observations to the strata under SRSS with the goal of achieving the most precise estimator possible. Finding such an optimal allocation is complicated because the variance of the SRSS estimator is not continuous in the number of observations. Thus, standard calculus-based optimization methods do not apply. The number of possible allocations is extremely large, so an exhaustive search takes considerable computation time. Attention is focused on methods that search subsets of the allocation space.

Simulated annealing is an attractive method because it is easy to program and computationally efficient.

---

*Yanqing Sun*, University of North Carolina at Charlotte, *Rajeshwari Sundaram*, National Institute of Child and Human Development, and *Yichuan Zhao*, Georgia State University

### **Empirical Likelihood Inference for the Cox Model with Time-Dependent Coefficients via Local Partial Likelihood**

The Cox model with time-dependent coefficients has been studied by a number of authors recently. In this paper, we develop empirical likelihood (EL) pointwise confidence regions for the time-dependent regression coefficients via local partial likelihood smoothing. The EL simultaneous confidence bands for a linear combination of the coefficients are also derived based on strong approximation methods. The empirical likelihood ratio is formulated through the local partial log-likelihood for the regression coefficient functions. Our numerical studies indicate that the EL pointwise/simultaneous confidence regions/bands have satisfactory finite sample performances. Compared with the confidence regions derived directly based on the asymptotic normal distribution of the local constant estimator, the EL confidence regions are overall tighter and can better capture the curvature of the underlying regression coefficient functions. Two data sets, the gastric cancer data and the Mayo Clinic primary biliary cirrhosis data, are analyzed using the proposed method.

---

*Zibonele Valdez-Jasso*, University of Texas at Dallas

### **Statistical analysis of UfMRI**

Functional Magnetic Resonance Imaging (fMRI) is a medical technique developed to scan brain activity when a patient is asked to perform a specific activity. Typically, each image is acquired every 1 or 2 seconds and is done over a period of time, providing a Time Series for each voxel. Researchers at UT Southwestern Medical Center developed a technique (UfMRI) which in general terms, selects only a slice of the brain and records images every .05 seconds. This allows observation of phenomena that could be wrongly classified as random noise when using traditional techniques. UfMRI also poses new statistical challenges; for example, a low signal to noise ratio of experimental responses. In this context, a wavelet denoising procedure, which is an aggregate of Universal and SureBlock methods, was developed. Theoretical properties will be presented, along with simulation results and examples using UfMRI data. (Joint work with *Sam Efromovich*)

---

*Haiyan Wang*, Kansas State University, *Siti Tolos*, Kansas State University, and *Suojin Wang*, Texas A&M University

## **A Nonparametric Test of Independence in the Presence of Heteroscedastic Treatment Effects**

In this paper, we present a test of independence between the response variable, which can be discrete or continuous, and a continuous covariate after adjusting for heteroscedastic treatment effects. The test statistic is constructed using moment methods after augmenting each pair of the data in each treatment with a fixed number of nearest neighbors as pseudo replicates. The asymptotic distribution of the proposed test statistic is obtained under the null and local alternatives. As a byproduct, we also give the test statistics and their asymptotic distributions for testing of no nonparametric effects of covariate and the treatment by covariate interaction. Though using a fixed number of nearest neighbors poses significant difficulty in the inference compared to that allowing the number of nearest neighbors to go to infinity, we obtained parametric standardizing rate for our test statistics in addition to reduced computational time. Numerical studies show that the new test procedures are powerful to detect nonlinear dependency.

---

*Suojin Wang*, Texas A&M University

## **Generalized Empirical Likelihood Methods for Analyzing Longitudinal Data**

Efficient estimation of parameters is a major objective in analyzing longitudinal data. In this work, we propose two generalized empirical likelihood based methods that take into consideration within-subject correlations. A nonparametric version of the Wilks theorem for the limiting distributions of the empirical likelihood ratios is derived. It is shown that one of the proposed methods is locally efficient among a class of within-subject variance covariance matrices. A simulation study is carried out to investigate the finite sample properties of the proposed methods and compare them with the block empirical likelihood method by You *et al.* (2006) and the normal approximation with a correctly estimated variance-covariance. The results suggest that the proposed methods are generally more efficient than existing methods which ignore the correlation structure, and better in coverage compared to the normal approximation with correctly specified within-subject correlation for small to moderate sample sizes. An application of the proposed procedures to the Framingham Heart Study is illustrated.

---

*Weihua Zhou*, University of North Carolina at Charlotte

## **A Multivariate Wilcoxon Regression Estimate**

An extension of univariate rank regression to multivariate linear models is proposed and studied. Unlike the coordinatewise rank regression considered by some earlier authors,

our approach is rotation equivariant, based on a multivariate spatial rank function introduced by Möttönen *et al.* (1995, 1997). The new estimate is unique, consistent and asymptotically normal under regularity conditions. The influence function of the estimate is derived and shown to be bounded in the  $y$  space but unbounded in the  $X$  space. Simulation shows that the new estimate is highly efficient if the errors of the linear model have a normal distribution and performs much better than the least squares estimate for heavy-tailed error distributions. The theory is illustrated with an example.

---