

### **Bandwidth Selection for Adaptive Density Estimation**

Mohamed Amezziane, Depaul University  
Tim McMurry, DePaul University

**Abstract:** We explore the usefulness of cross validation as a means of smoothing parameter selection choice for density estimators which adapt to the smoothness of the unknown density. Particular attention will be paid to kernels of the infinite order at-top class. Infinite order estimators offer potentially rapid rates of convergence, and cross validation is often an attractive method of bandwidth selection requiring minimal assumptions. We quantify the relationship between the performance of cross validation, the kernel, and smoothness characteristics of the unknown density, and discuss the conditions under which cross validation is a potentially useful approach. Theoretical results are bolstered by an extensive simulation study of multimodal and non-smooth functions.

---

### **A Comparison of Frequentist and Bayesian Approaches for Linear Gaussian Process Models**

Muhammad Atiyat, Penn State University  
Murali Haran, Penn State University

**Abstract:** Spatial data (data that are geographically referenced) are commonly encountered in varied fields such as ecology, epidemiology, public health, and geoscience. We consider practical issues with using linear Gaussian process models, which are among the most popular models for analyzing spatial data. We summarize some commonly used frequentist and Bayesian approaches for modeling spatial data via Gaussian processes. In the Bayesian context we review some standard approaches for selecting appropriate priors. We also compare estimation and prediction for Gaussian process models via a simulation study and through an application of our methods to a spatial data set used for studying crop epidemics. We conclude with some practical recommendations based on our study.

---

### **Nonparametric Estimation in Multivariate Mixture Models**

Tatiana Benaglia, Penn State University  
Didier Chauveau, Université d'Orléans, France  
David R. Hunter, Penn State University

**Abstract:** We propose an algorithm for nonparametric estimation for finite mixtures of multivariate random vectors that is not, but that strongly resembles, a true EM algorithm. The vectors are assumed to have independent coordinates conditional upon knowing which mixture component from which they come, but otherwise their density functions are completely unspecified. Sometimes, the density functions may be partially specified by Euclidean parameters, a case we call semiparametric. Our algorithm is much more flexible and easily applicable than existing algorithms in the literature; it can be extended to any number of mixture components and any number of coordinates of the multivariate observations. Thus it may be applied even in situations where the model is not identifiable, so care is called when it is difficult to establish identifiability conclusively. Our algorithm yields much smaller mean integrated squared errors than an alternative algorithm in a simulation study. In another example using a real dataset, it provides new insights that extend previous analyses.

### Combining Complex Climate Models with Massive Observational Data for Predicting Climate Change

K. Sham Bhat, Penn State University  
Murali Haran, Penn State University  
Klaus Keller, Penn State University  
Joshua Dorin, Penn State University

**Abstract:** We study the risk of a collapse of the meridional overturning circulation (MOC), part of the global ocean circulation “conveyor belt”, a key factor determining global climate patterns. Assessing this risk involves combining massive computer climate model output with space-time intensive physical observations of the ocean system. We use a Bayesian approach to connect the two sets of data by approximating the climate model output using a Gaussian process based emulator and using Markov Chain Monte Carlo (MCMC) methods to obtain a posterior distribution for important parameters associated with the probability of an MOC collapse. We use a kernel convolutions approach (Higdon 1998) to make our approach computationally tractable.

---

### Powerful Nonparametric Tests in Shape Analysis for Large Number of Variables and Few Subjects

Chiara Brombin, University of Padova, Italy  
Fortunato Pesarin, University of Padova, Italy  
Luigi Salmaso, University of Padova, Italy

**Abstract:** Inferential methods in shape analysis make use of configurations of landmarks optimally superimposed using a least-squares procedure or analyzing matrices of interlandmark distances. In the two independent samples case, the most natural way to compare two mean shapes is by means of the well-known Hotelling’s  $T^2$  test. Despite its widespread use, it is well known that Hotelling’s  $T^2$ -test may not be very powerful unless there are a large number of observations available (Dryden and Mardia, 1998 and Blair et al., 1994), but very frequently researchers have to cope with few individuals and many landmarks. For these reason we propose a  $T^2$ -type test in a nonparametric permutation framework within the nonparametric combination approach (Pesarin, 2001). A comparative simulation study, under the multivariate normal distribution, will show that the power for the suggested test increases when increasing the number of the processed variables provided that the noncentrality parameter  $\delta$  increases, even when the number of covariates is larger than the permutation sample space. Similar results have been obtained under other distributions, like Cauchy, Student’s  $t$  with 2 d.f. and Pareto (with shape and shape-scale parameters). Such findings are very important in order to carry out a shape analysis even in case of small sample sizes and a large number of landmarks or semi-landmarks.

### Statistical Models for Globular Cluster Luminosity Distribution

Max Buot, Xavier University  
Donald Richards, Penn State University

**Abstract:** We consider statistical models which have been proposed for luminosity distributions for the globular clusters in the Milky Way and M31. Although earlier research showed that the cluster luminosity functions in those two galaxies were well fit by Gaussian distributions, subsequent investigations suggested that their luminosities followed t-distributions rather than Gaussian distributions. By applying the Bayesian Information Criterion, we find moderate statistical evidence that the t-distribution is superior to the Gaussian distribution as a model of luminosity distribution for the Milky Way. In the case of M31, we find moderate evidence that the Gaussian distribution is superior to the t-distribution. Our conclusion is that in neither case do we find strong evidence to support the use of one distribution over the other as a statistical model for the luminosities of the globular clusters in the Galaxy and M31. Moreover, we urge caution in the use of the Kolmogorov-Smirnov statistic to justify the choice of statistical models for globular cluster luminosity functions.

---

### The Likelihood-Tuned Density Estimator

Yejin Chung, Penn State University  
Bruce G. Lindsay, Penn State University

**Abstract:** We consider an improved density estimator which arises from treating the kernel density estimator as an element of the model that consists of all mixtures of the kernel, continuous or discrete. One can then “likelihood tune” the kernel density estimator by using it appropriately as the starting value in an EM algorithm. If we do so, then one EM step leads to a fitted density with higher likelihood than the kernel density estimator. The one step EM estimator can be written explicitly, and its bias is one order of magnitude smaller than the kernel estimator. In addition, the order of magnitude of the variance stays of the same order, so that the asymptotic mean square error can be reduced significantly. Compared with other important adaptive density estimators, we find that their biases are in the same order but our estimator is still superior, particularly when the density is small. We also compare the mean squared error of this new density estimator with pre-existing estimators using simulation results.

---

### Monte Carlo Study on the Performance of AIC and BIC in Latent Class Analyses with Various Sample Sizes

John Dziak, Penn State University

**Abstract:** The effect of sample size on the correct model size selection rate and on parameter estimation error in latent class analysis with AIC or BIC model selection criteria was explored in a realistic simulation example. The varying, but generally bad, performance of both classical information criteria suggest some of the difficulties in finding, or even defining, a correct model size in practice, as well as some difficulties in interpreting Monte Carlo studies of model selection.

### Geometry and Robustness in Basic Statistics

Ryan T. Elmore, Colorado State University  
Paul W. Mielke, Jr., Colorado State University

**Abstract:** It is well known that the one-way analysis-of-variance (ANOVA)  $F$  test and the two-sample  $t$  test are very intolerant to the existence of a relatively few extreme values. In this paper, we will discuss the underlying geometry of these methods and how it relates to this problem with outliers. The robust alternative based on ranks (Wilcoxon-Mann-Whitney) will be examined from a geometric perspective as well. Because these classical statistical methods belong to a broad class of statistical tests termed multi-response permutation procedures (MRPP) which are based on distance functions, both the problem and an alternative solution may be simply described in the context of MRPP.

---

### Nonparametric Conditional Quantile Estimation with Neural Network

Yijia Feng, Penn State University

**Abstract:** We study the robust neural network (RNN) for nonparametric conditional quantile estimation. We proposed an MM algorithm to realize the optimization of the general quantile loss. A simulation study is performed to compare the proposed RNN with some other nonparametric regression methods. An application of our method is presented to estimate the maturation curve in a credit card portfolio dataset.

---

### Some Properties of Three Two-Sample Tests

Roger L. Goodwin, US Government Printing Office

**Abstract:** Using the Neyman-Pearson Lemma, we will develop some best critical regions for three two-sample dispersion tests. A source of confounding occurs in these distribution free tests — namely the sample sizes  $n$  and  $m$ . When the sample sizes are not equal, then we see that distribution free tests become more of a test of which sample is larger. For large differences between the two sample sizes, this becomes a problem. The best critical region of the Miller jackknife test is based upon the  $\log \chi^2$  distribution. However, much of the literature quotes the normal distribution as being an approximation for this test even though the  $\ln$  function of the variances are clearly calculated during the jackknife procedure. Based on statistical power calculations and analyses, the Savage exponential test is superior to both the Miller jackknife test and the Capon-Klotz test. However, the Savage test for dispersion essentially tests for the differences in the means. But, the sufficient statistics for the means are the sufficient statistics for the variance parameter in the exponential distribution. Under a given set of conditions, it will be shown that the Miller test is superior over the Capon-Klotz test and vice-versa.

### Nonparametric Regression to the Mean via Kernel Methods

Majnu John, Children's Hospital of Philadelphia  
Abbas F. Jawad, University of Pennsylvania

**Abstract:** Part of the estimated longitudinal change over time of a response variable could be explained by the regression to the mean phenomenon. The component of change due to regression to the mean is more pronounced with subjects whose initial measurement tends to be extreme. Das and Mulder (1983) proposed a nonparametric approach to estimate the regression to the mean. Data-adaptation for empirical distributions via kernel estimation methods is considered in this poster, which utilize Das and Mulder's method and their original assumptions. The best kernel methods for density and hazard function estimation were used. This makes our approach extremely user friendly for practitioners via the state of the art procedures and packages available in statistical software such as SAS and R for kernel density and hazard function estimation. We estimate the standard error of regression to the mean via nonparametric bootstrap methods. We also conduct a simulation study to compare the methods for different underlying densities. Our estimation approach is particularly suited for many response variables in clinical research such as, for example, FEV1%, blood pressure, Hba1c, and many other lipid measurements.

---

### Local Polynomial Composite Quantile Regression

Bo Kai, Penn State University  
Runze Li, Penn State University  
Hui Zou, University of Minnesota

**Abstract:** Nonparametric regression is a useful statistical tool to explore fine features in the data, and has been applied in various disciplines. In this presentation, we propose local polynomial composite quantile regression for nonparametric regression models. We derive the asymptotic bias, variance and normality of the proposed estimate. Asymptotic relative efficiency of the proposed estimate to the local polynomial regression under the least squares loss is investigated. We show that the proposed estimate can be much more efficient than the local polynomial regression estimate with the squared loss for various non normal errors, and is almost as efficient as the LS estimate for normal error. Simulation is conducted to examine the performance of the proposed estimates. The simulation results are consistent with our theoretic findings. A real data example is used to illustrate the proposed procedures.

---

### Recent History of Functional Linear Models

Kion Kim, Penn State University  
Damla Senturk, Penn State University

**Abstract:** We propose a variant of historical functional linear models for cases where the current response is affected by the predictor process in a window into the past. Different from the rectangular support of functional linear models, the triangular support of the historical functional linear models and the point-wise support of the varying coefficient models, the current model has a sliding window support into the past. This idea leads to models that bridge the gap between varying coefficient models and functional linear (historic) models. The proposed estimation algorithm is shown to be fast in comparison to estimation procedures proposed for historical functional linear models, involving one dimensional basis expansions and one dimensional smoothing procedures.

### Location Tests in the IC Model Using Marginal Ranks

Klaus Nordhausen, University of Tampere, Finland  
Hannu Oja, Tampere School of Public Health, Finland  
Davy Paindaveine, Université Libre de Bruxelles, Belgium

**Abstract:** A common multivariate model formulation is

$$X_i = \lambda Z_i + \mu, i = 1, \dots, n,$$

where  $Z_i$  is centered at the origin. For inference about the location parameter  $\mu$  the standard parametric test is Hotelling's  $T^2$  which assumes that  $Z_i$  is normal distributed or has at least finite second order moments. A fully nonparametric counterproposal can be based on the marginal signs and ranks of  $X_i$  as for example described in Puri and Sen (1971). These nonparametric location tests are however unfortunately not affine equivariant. The tests introduced here will also use the marginal signs and ranks, however not those of  $X_i$  but of  $Z_i$ , assuming the components of  $Z_i$  are independent and symmetric. Under these assumptions we assume the data follows the so called restricted independent component model and the first step needed is to recover the unobserved values of  $Z_i$ , which is normally done by estimating  $\lambda$ . The tests introduced here need only a  $\sqrt{n}$ -consistent estimate of  $\lambda$  that is not effected under individual sign changes of observations.

We will show asymptotic as well as finite sample efficiencies of the test using different score functions compared to Hotelling's  $T^2$  and compare the robustness of the tests when outliers in  $Z_i$  are present. The results shown are obtained by applying the two different scatter matrices method of Oja et al. (2006) to estimate  $\lambda$  where moment assumptions can be avoided.

---

### Nonparametric Tests of Circularity of Complex-Valued Random Vectors

Esa Ollila, University of Oulu, Finland  
Visa Koivunen, Helsinki University of Technology, Finland

**Abstract:** An important characteristic of a complex-valued random vector is the so called circularity property: a complex random vector is said to be circular (or circularly symmetric) if its distributions remains invariant under multiplication by unit-modulus complex scalar. In this paper, we consider likelihood ratio and nonparametric tests of circularity of complex random vectors. The developed nonparametric tests are based on the concepts of multivariate complex-valued spatial sign and rank vectors and complex-valued equivalent of Tyler's M-estimator of scatter.

### Adaptive Nonparametric Likelihood Weights and Mixtures of Empirical Distributions

Jean-François Plante, University of Toronto  
James V. Zidek, University of British Columbia

**Abstract:** Suppose that you must infer about a population, but that data from  $m-1$  similar populations are available. The weighted likelihood uses exponential weights to include all the available information into the inference. The contribution of each datum is discounted based on its dissimilarity with the target distribution.

We propose a nonparametric method to determine likelihood weights based on the data. The suggested “MAMSE” weights can be used as likelihood weights, or as mixing probabilities to define a mixture of empirical distributions.

The MAMSE weights are defined for different types of data: univariate, right-censored and multivariate. In addition to their role for the likelihood, the MAMSE weights are used to define a weighted Kaplan-Meier estimate of the survival distribution and weighted coefficients of correlation based on ranks. The maximum weighted pseudo-likelihood, a new extension of a method used to fit families of copulas, is also proposed. All these examples of inference using the MAMSE weights are shown to be consistent. Furthermore, simulations show that inference based on MAMSE-weighted methods can perform better than their unweighted counterparts. Hence, the nonparametric adaptive weights we propose successfully borrow strength from the  $m - 1$  similar populations whose distribution may differ from the target.

---

### Consistency and Asymptotic Distribution of the Theil-Sen Estimator

Shaoli Wang, Yale University

**Abstract:** In this paper, we obtain the strong consistency and asymptotic distribution of the Theil-Sen estimator in simple linear regression models with arbitrary error distributions. We show that the Theil-Sen estimator is super-efficient when the error distribution is discontinuous and that its asymptotic distribution may or may not be normal when the error distribution is continuous. We give an example in which the Theil-Sen estimator is not asymptotically normal. A small simulation study is conducted to confirm the super-efficiency and the non-normality of the asymptotic distribution.

---

### Semi-Parametric Estimation for Repeated Measures in Finite Mixture Models

Tracey Wrobel, Penn State University  
Thomas P. Hettmansperger, Penn State University

**Abstract:** Methods are developed to handle repeated measures mixture models. Prior work in repeated measures mixtures assumed that, given component membership, the repeated measures on a subject are independent and identically distributed. The method we propose requires only that the repeated measures be independent given component membership, not necessary identically distributed. Our method assumes that the component distributions are related by an exponential tilt. This is a semi-parametric approach since, other than the exponential tilt, the marginal distributions of the observations are unspecified. This method is an extension of unpublished work by Jing Qin and Denis Leung and includes a profile empirical likelihood and an EM algorithm to estimate the parameters in the exponential part.

### Regularization Parameter Selection for Penalized Likelihood Functions

Michael Yiyun Zhang, Penn State University

**Abstract:** We apply the nonconcave penalized likelihood approach to obtain variable selections as well as shrinkage estimators. This approach relies heavily on the choice of regularization parameter, which controls the model complexity. We propose employing the generalized information criterion (GIC) for selecting the regularization parameter, and study the asymptotic behavior of several types of GIC selectors. The results bridge the classical variable selection criteria and the penalized methods. Our simulation results confirm the theoretical findings.