Understanding of the New Statistics: Expanding Core Statistical Theory

Rudolf Beran (University of California Davis)
Iain Johnstone (Stanford University)
Ivan Mizera (University of Alberta)
Sara van de Geer (ETH Zürich)

September 14, 2008 – September 19, 2008

1 Overview of the Field, and Recent Developments

As proposed, the focus of the workshop was core statistical theory—something that in past might have been called "mathematical statistics", were not this name so closely associated with the 1960's vision of statistics before the computer revolution changed the discipline. The emphasis was on "the new statistics", relevant theory for the data-analytic circumstances. (The note [1], authored by one of the organizers, brings more detailed historical analysis and discussion of recent aspects.)

Several participants of the workshop recognized somewhat polemical nature of the term "new statistics", which quickly became a sort of recurrent mantra in the discussions. However, "new" does not imply negation of "old" here. The viewpoint of one of the former editors of JASA, that the theoretical underpinnings of statistical science "hold firm as ever", was not really challenged at the meeting. While the past have seen a lot of crystallization of ideas one one hand and fierce ideological disputes on another, the present is characterized not any of these prevailing over others, but by a totally different, in mathematical parlance orthogonal angle of view.

Such a perception shift was caused in the first place by the burgeoning development of "statistical sciences", that is, disciplines that rely on statistical theory to express their specific ideas, like econometrics, genomics, signal processing, or machine learning. These days, statisticians can be found at various places, not necessarily mathematical or statistical departments; for instance, among Nobel laureates in economics. The turmoil in statistical sciences led to a tremendous reevaluation of the "core of statistics" as defined in [3]: the subset of statistical activity that is focused inward, on the subject itself, rather than outward, towards the needs of statistics in particular scientific domains. The character of this reappraisal did not have that much character of confirmation or refutation, but more that of assessing relevance and usability. Although nothing is definitive in this respect—the supposedly barren fields of yesterday may turn fertile irrigated by the demands of applications—statisticians witnessed that some of the items on their buffet tables were in strong demand, and some of them remained unnoticed. As emphasized in the talks of several participants, the past differences in tastes (say, regarding the use or abuse of inverse probability in statistical deliberations) are felt to be much less divisive now; there are issues now that are more important than sectarian quarrels.

To acquire some feeling about the current situation, the workshop brought together a number of leading-edge researchers in mathematical statistics (a considerable number of them affiliated with computer science rather than classical statistical departments), together with key people from the statistical sciences and other communities practicing data analysis—to assess, in certain selected areas (hopes to seize such an important and complex problematic in its totality would be indeed futile), "what is hot and what is not". The results were partly expected, partly surprising. The areas of concentration were the following.

2 Presentation Highlights

2.1 The New Asymptotics and Random Matrices

One of the possible approaches capable of formalizing some phenomena occurring in "large p, small n", is the asymptotic theory of random matrices, the theme of the opening talk of the workshop, Null Distributions for Largest Eigenvalues in Multivariate Analysis, presented by Iain Johnstone (Stanford). The theme was eigenvalues of Wishart matrices: while these play a central role in classical multivariate analysis, a new impetus to approximate distribution results has come from methods that imagine the number of variables as large. Johnstone focused on the largest eigenvalue in particular, and briefly reviewed null distribution approximations in terms of the Tracy-Widom laws. The second part of Johnstone's talk described the work in progress on concentration inequalities for the largest eigenvalue in the "two Wishart" case, such as canonical correlations.

Noureddine El Karoui (Berkeley) in Spectral Properties of Kernel Matrices with High-Dimensional Input Data addressed kernel matrices, a special form of matrices used in various areas of machine learning and statistics. For instance, they are sometime used to perform non-linear versions of principal component analysis (PCA). There has been little work so far investigating their spectral properties when the dimension of the data (p) is of the same order of magnitude as the number of observations (n). El Karoui discussed some results concerning this asymptotic setting, assuming standard and less standard models used in random matrix theory for the data. In particular, it can be shown that for these models, kernel random matrices behave essentially like linear operator—a sharp difference with the low dimensional setting where p is held fixed—where they approximate the spectra of certain integral operators. El Karoui discussed also the proof, highlighting some potential geometric limitations of standard random matrix models for statistical applications, and also some robustness (and lack thereof) results against these geometric features for classical random matrix results.

Random matrix theory then became one of the recurring themes in the other days. Art B. Owen (Stanford, jointly with Patrick Perry) presented a talk entitled Cross-Validation for the Truncated SVD and the Non-Negative Matrix Factorization. Owen and Perry study sample reuse methods like the bootstrap and cross-validation, methods which are widely used in statistics and machine learning. These methods provide measures of accuracy with some face value validity that is not dependent on strong model assumptions. They depend on repeating or omitting cases, while keeping all the variables in those cases. But for many data sets, it is not obvious whether the rows are cases and columns are variables, or vice versa. For example, with movie ratings organized by movie and customer, both movie and customer IDs can be thought of as variables. Owen and Perry looked at bootstrap and cross-validation methods that treat rows and columns of the matrix symmetrically. They got the same answer on X as on X'. While McCullagh has proved that no exact bootstrap exists in a certain framework of this type (crossed random effects), they showed that a method based on resampling both rows and columns of the data matrix tracks the true error, for some simple statistics applied to large data matrices. They also looked at a method of cross-validation that leaves out blocks of the data matrix, generalizing a proposal due to Gabriel that is used in the crop science literature. They found empirically that this approach provides a good way to choose the number of terms in a truncated SVD model or a non-negative matrix factorization and applied some recent results in random matrix theory to the truncated SVD case.

The connection to the asymptotics of random matrices was imminent also in the topic of the talk of Marc Hallin (Universiteé Libre de Bruxelles) who spoke about The General Dynamic Factor Model developed by Forni et al. (2000) for the analysis of large panels of time series data. Although developed in an econometric context, this method is likely to apply in all fields where a very large number of interrelated time series or signals are observed simultaneously. Hallin considered the problem of identifying the number q of factors driving the panel. The proposed criterion was based on the fact that q is also the number of diverging eigenvalues of the spectral density matrix of the observations as the cross-sectional dimension n goes to infinity. Hallin gave sufficient conditions for consistency of the criterion for large n and T (where T is the series length), showed how the method can be implemented, and presented simulations and empirics illustrating its excellent finite sample performance. The application to real data brought some new contribution in the debate on the number of factors driving the US economy.

2.2 The New Asymptotics and Sparse Representations

Sara van de Geer (ETH Zürich, jointly with Peter Bühlmann and Lukas Meier) in her talk on High-Dimensional Additive Modeling considered a high-dimensional additive regression model of the form

$$y = f_1(x) + \cdots + f_p(x) + \epsilon,$$

where the f_j are unknown smooth functions of the covariable x, and where p is large, possibly larger than the number of observations n. She proposed to use penalized least squares to estimate the f_j. The first question to address is then: how to choose the penalty? For this purpose, she first considered the penalized least squares problem in general terms, with splines and the lasso as special cases; then she discussed the trade-off between theoretically 'ideal' but computationally less ideal penalties. With these insights, it is possible to consider 3 possible penalties for the high-dimensional additive model, having in mind the "sparse" situation, where most of the f_j are actually zero. She presented results amounting to oracle type of inequalities in this setting, a simulation study, and an application to real data.

Sparsity was then the theme of several ensuing talks. Marten Wegkamp (Florida State) in Generalized Support Vector Machines in Sparse Settings considered the problem of binary classification where one can, for a particular cost, choose not to classify an observation, and presented a simple oracle inequality for the excess risk of structural risk minimizers, using a generalized hinge loss and a lasso type penalty. He showed that it is possible to obtain fast rates, regardless of the behavior of the conditional class probabilities, if the Bayes discriminant function can be well approximated by a sparse representation.

Florentina Bunea (Florida State) in Honest Inference in Sparse High-Dimensional Models studied the topic of ℓ_1 regularized or lasso type estimation, the topic that has received considerable attention over the past decade. Recent theoretical advances have been mainly concerned with the risk of the estimators and corresponding sparsity oracle inequalities. In her talk, Bunea investigated the quality of the ℓ_1 (and the closely related $\ell_1 + \ell_2$) penalized estimators from a different perspective, shifting the emphasis to correct, non-asymptotic, model identification. She illustrated the merits and limitations of these methods in high dimensional generalized regression and mixture models and discussed an important consequence of this analysis: if the "true" model is identifiable, good estimation/prediction properties of the parameter estimates are not necessarily needed for correct model identification.

An small sensation of the first day was Benedikt M. Pötscher (Vienna), who in his talk, Confidence Sets Based on Sparse Estimators Are Necessarily Large, pointed out some deficiencies in some of the recent sparsity results, deficiencies that can be seen as special cases of some classical and well-known, but perhaps somewhat forgotten phenomena. In particular, he showed that the confidence sets based on sparse estimators are large compared to more standard confidence sets, which demonstrates that the sparsity of an estimator comes at a substantial price in terms of the quality of the estimator. His results were set in a general parametric or semiparametric framework.

A different approach to sparse representation was presented by Lawrence Brown (University of Pennsylvania, jointly with Eitan Greenstein). In Non-parametric Empirical Bayes and Compound Bayes Estimation of Independent Normal Means, Brown considered the classical problem of estimating a vector $\mu = (\mu_1, \ldots, \mu_n)$, based on independent observations $Y_i \sim N(\mu, \sigma^2)$, where μ_i themselves are independent realizations from a completely unknown distribution G_n. Brown proposed an easily computed estimator $\tilde{\mu}$ and studied the ratio of its expected risk $E_{G_n} E_\mu(\|\tilde{\mu} - \mu\|)$ to that of the Bayes procedure. He showed that under very mild conditions, this ratio approaches 1 as $n \to \infty$, and considered also a related compound decision theoretic formulation, where this estimator is asymptotically optimal relative to the best possible estimator given the values of the order statistics $\mu_{(*),n} = (\mu_{(1)}, \ldots, \mu_{(n)})$. In the discussion of this proposal, Brown reminded the audience that there has been much contemporary interest in estimators that are valid in sparse settings; settings such as those for which $p_n \to 0$ and $G_n(\{u\}) = p_n$ if $u = 0, = 1 - p_n$ if $u = u_n$. The conditions on the sequences G_n or $\{\mu_{(*),n}\}$ for asymptotic optimality of $\tilde{\mu}$ are only mildly restrictive, and include a broad range of problems involving sparsity. In particular, the proposed estimator is asymptotically optimal in moderately "sparse" settings - ones such as those described just above in which $np_n \to \infty$

and $n(1-p_n) \to \infty$ $0 < \liminf u_n, \limsup u_n < \infty$. He also reported a simulation study to demonstrate the performance of our estimator, showing that in moderately sparse settings his estimator performs very well in comparison with current procedures tailored for sparse situation, and adapts also well to non-sparse situations, and concluded his talk by a very interesting application to baseball data.

Kjell Doksum (University of Wisconsin, jointly with Shijie Tang and Kam Tsui) focused On Nonparametric Variable Selection, in the setting of regression experiments involving a response variable Y and a large number d of predictor variables (X's) many of which may be of no value for the prediction of Y and thus need to be removed before predicting Y from the X's. His talk considered procedures that select variables by using importance scores that measure the strength of the relationship between predictor variables and a response. In the first of these procedures, scores are obtained by randomly drawing subregions (tubes) of the covariate space that constrain all but one predictor and in each subregion computing a signal to noise ratio (efficacy) based on a nonparametric univariate regression of Y on the unconstrained variable. The regions are adapted to boost weak variables iteratively by searching (hunting) for the regions where the efficacy is maximized. The efficacy can be viewed as an approximation to a one-to-one function of the probability of identifying features. By using importance scores based on averages of maximized efficacies, we develop a variable selection algorithm called EARTH (Efficacy adaptive regression tube hunting). The second importance score method (RFVS) is based on using Random Forest importance values to select variables. Computer simulations show that EARTH and RFVS are successful variable selection methods when compared to other procedures in nonparametric situations with a large number of irrelevant predictor variables. Moreover, when each is combined with the model selection and prediction procedure MARS, the tree-based prediction procedure GUIDE, or the Random Forest prediction method, the combinations lead to improved prediction accuracy for certain models with many irrelevant variables. Doksum gave conditions under which a version of the EARTH algorithm selects the correct model with probability tending to one as the sample size tends to infinity, even if d tends to infinity as n tends to infinity, and concluded his talk with the analysis of a real data set.

In a similar vein, Hannes Leeb (Yale) proposed methodology for Evaluating and Selecting Models for Prediction Out-Of-Sample. The problem he studied, in the framework of the regression with random design, was that of selecting a model that performs well for out-of-sample prediction, focusing on a statistically challenging scenario where the number of potentially important explanatory variables can be infinite, where no regularity conditions are imposed on unknown parameters, where the number of explanatory variables in a "good" model can be of the same order as sample size, and where the number of candidate models can be of larger order than sample size.

2.3 The Role of Geometry

The geometric line of the workshop started by the talk of Peter Kim (Guelph, jointly with Peter Bubenik, Gunnar Carlsson, and Zhiming Luo) entitled Geometric and Topological Methods for High-Dimensional Data Analysis. Kim examined the estimation of a signal embedded in white noise on a compact manifold, where a sharp asymptotic minimax bound can be determined under the sup-norm risk over Hölder classes of functions, generalizing similar results available for spheres in various dimensions. The estimation allows for the development of a statistical Morse theory using the level sets of the estimated function and together with the sup-norm bound allows the bounding of the Hausdorff distance in a persistence diagram in computational algebraic topology.

Mikhail Belkin (Ohio State) in his talk Spectral and Geometric Methods in Learning discussed some of the methods from a variety of spectral and geometry-based methods that became popular for various tasks of machine learning (such as dimensionality reduction, clustering and semi-supervised learning) and recent theoretical results on their convergence. A particularly interesting part of his talk was a proposal how spectral methods can be used to estimate parameters in the mixtures of Gaussian distributions.

The correspondence with information-theoretical aspects through convex geometry embedded in the duality theory of convex optimization was the theme of two talks. Michael Jordan (Berkeley, jointly with XuanLong Nguyen and Martin Wainwright), in On Surrogate Loss Functions and f-Divergences) looked at binary classification, where the goal is to estimate a discriminant function γ from observations of covariate vectors and corresponding binary labels. He considered an elaboration of this problem in which the covariates are not available directly, but are transformed by a dimensionality-reducing quantizer Q, and presented conditions on loss functions such that empirical risk minimization yields Bayes consistency when both the discriminant function and the quantizer are estimated. These conditions were stated in terms of a general correspondence between loss functions and a class of functionals known as Ali-Silvey or f-divergence functionals. Whereas this correspondence was established by Blackwell (1951) for the 0-1 loss, Jordan presented an extension of the correspondence to the broader class of surrogate loss functions that play a key role in the general theory of Bayes consistency for binary classification. The result makes it possible to pick out the (strict) subset of surrogate loss functions that yield Bayes consistency for joint estimation of the discriminant function and the quantizer.

Ivan Mizera (University of Alberta, jointly with Roger Koenker) in Quasi-Concave Density Estimation: Duality in Action explored the duality aspects in maximum likelihood estimation of a log-concave probability density formulated as a convex optimization problem. It was shown that an equivalent dual formulation is a constrained maximum Shannon entropy problem. Mizera considered also closely related maximum Renyi entropy estimators that impose weaker concavity restrictions on the fitted density, notably a minimum Hellinger discrepancy estimator that constrains the reciprocal of the square-root of the density to be concave; a limiting form of these estimators constrains solutions to the class of quasi-concave densities.

2.4 Algorithmic Inspirations

Speakers in this area mostly addressed various aspects of regularization as recently used in statistical methods. Saharon Rosset (Tel Aviv, jointly with G. Swirzscz, N. Srebro, J. Zhu) in ℓ_1 Regularization in Infinite Dimensional Feature Spaces discussed the problem of fitting L1 regularized prediction models in infinite (possibly non-countable) dimensional feature spaces. The main contributions were: (a) Deriving a generalization of L1 regularization based on measures which can be applied in non-countable feature spaces; (b) Proving that the sparsity property of L1 regularization is maintained in infinite dimensions; (c) Devising a path-following algorithm that can generate the set of regularized solutions in ”nice” feature spaces; and (d) Presenting an example of penalized spline models where this path following algorithm is computationally feasible, and gives encouraging empirical results.

Roger Koenker (University of Illinois) in Computational Pathways for Regularized Quantile Regression recalled that in the beginning was the weighted median of Boscovich and Laplace; much later, Edgeworth (1888) nearly discovered the simplex algorithm, and Frisch (1956) almost proposed the interior point (log-barrier) method for linear programming; finally, modern variants of these methods, aided by recent developments in sparse linear algebra, are highly effective in many statistical applications. However, in regression settings with large, dense designs these methods perform quite poorly. Fortunately, old-fashioned simplex-type parametric programming methods come to the rescue for some problems of this type. Koenker reported on some computational experience in such situations, and make some more speculative remarks on implications for the choice of the ubiquitous regularization parameter. A common feature of the foregoing approaches to computation is that they all seek the path to enlightenment via some form of regularization.

The algorithmic inspiration of Giles Hooker (Cornell) for his Inference from Black Boxes came from the fields of machine learning or data mining, which produced a multitude of tools for “algorithmic learning”. Such tools are frequently ad hoc in nature, justified by some heuristics and can be challenging to analyze mathematically. Moreover, these routines produce prediction prediction functions that are typically algebraically complex and difficult to interpret.Nonetheless, there has been considerable interest in tools to “x-ray the black box”. Many of these tools can be understood in terms of the functional ANOVA decomposition. This decomposition represents a high dimensional function in terms of an additive expansion of lower dimensional components and allows us to quantify measures like variable importance and the average effect of certain variables. In his talk, Hooker examined the practical and theoretical challenges in turning such diagnostic procedures into formalized statistical tests. Specifically, he examined bootstrap inference about the functional ANOVA relationships in the underlying structure of the data. His aim was to be generic in the sense of providing tools that are universally applicable, regardless of the learning algorithm employed.

A yet another application of regularization methods, Inpendence and Conditional Independence with Reproducing Kernels, was presented by Kenji Fukumizu (Institute of Statistical Mathematics). He proposed new nonparametric methodology for dependence of random variables, with application to dimension reduction for regression. The methodology uses the framework of reproducing kernel Hilbert spaces (RKHS) defined by positive definite kernels. In this methodology, a random variable is mapped to a RKHS, thus random variables on the RKHS are considered. Fukumizu showed that the basic statistics such as mean and covariance of the variables on the RKHS can capture all the information on the underlying probabilities, and provide a method of characterizing independence and conditional independence. The framework of RKHS enables to derive a practical and efficient way of computing estimators defined on RKHS. Using the characterization of conditional independence, Fukumizu introduced a method of dimension reduction or feature extraction of the covariates is introduced for regression problems, and derived a practical algorithm to extract an effective linear feature. The method is of wide applicability; it does not require any strong assumptions on the type of variables or the probability of variables, which are often imposed by other methods of dimension reduction. Consistency of the estimator was proved under weak condition, and some experimental results show the method is practically competitive.

Rudolf Beran (Davis) in Penalized Fits of Multivariate Responses to Covariates considered a complete k-way layout of d-dimensional mean vectors, in which each mean vector is an unknown function of k real-valued covariates whose values are known, and the covariates may be either ordinal or nominal. There is at least one observation with error on each of the unknown mean vectors; the problem is to estimate the mean vectors efficiently, without making any assumption about the function that relates them to the known covariates. Both theory and practice have made it clear that the unconstrained least squares estimator of the mean vectors is unsatisfactory unless the data provides substantial replication. In his talk, Beran defined a candidate class of penalized least squares (PLS) estimators suitable for the problem. A separate quadratic penalty term is devised for each of the main effects and interactions in the MANOVA decomposition of the mean vectors. The construction of the penalty terms draws on vague notions about the unknown function that links the means to the covariates. Before being summed, each penalty term is weighted by right multiplication with a $d \times d$ symmetric positive semidefinite matrix. The candidate PLS estimators thereby accomplish, as special cases, both MANOVA submodel selection and dimensionality reduction. The matrix penalty weights were chosen to minimize estimated quadratic risk over the candidate class of PLS estimators; it was shown that, as the number of cells in the complete k-way layout tends to infinity, the candidate PLS estimator with smallest estimated risk converges, in loss or risk, to the candidate estimator with smallest actual loss or risk. The asymptotics make no assumptions about the unknown d-dimensional mean vectors in the k-way layout and require no replication. The talk was concluded by a case study on multivariate response data illustrating how the proposed adaptive estimator works.

A different, but very appealing inspiration by algorithmic methods was presented by Marloes Maathuis (ETH Zürich, jointly with Markus Kalisch and Peter Bühlmann), who studied Variable Importance Based on Intervention Calculus. It is assumed that we have observational data, generated from an unknown underlying directed acyclic graph (DAG) model, and it is well-known that a DAG is not identifiable from observational data, but it is possible to consistently estimate an equivalence class of DAGs. Moreover, for any given DAG, causal effects can be estimated using intervention calculus. In her talk, Maathuis combined these two parts. For each DAG in the estimated equivalence class, she used intervention calculus to determine the causal effects of the covariates on the response. This yields a collection of possible causal effects for each covariate. Maathuis showed that the distinct values in this set can be consistently estimated by an algorithm that uses only local information of the graph. This local approach is computationally fast and also has advantages from an estimation point of view. Maathuis proposed to use summary measures of the set of possible causal effects to determine variable importance; in particular, to use the minimum absolute value of this set, since that is a conservative bound on the size of the causal effect.

2.5 Old as a Backbone of the New Statistics

Two of the talks happened to be aimed at a reflection about foundations. Bertrand Clarke (University of British Columbia) proposed Coordinating Theory, motivated by the impact of new data types and the compartmentalization of subfields of statistics. His theory intends to interrelate the disparate principles and practices of Statistics within one framework in, as he argued, “a good time to crystallize what unites us in statistics.” The key features are predictive optimality and a unified variance bias treatment; the approach includes Bayes, Frequentist and other perspectives, including subsidiary criteria, such as robustness and efficiency. An an application of this framework, Clarke presented a comparison of three predictors in the context of a complex data set, formalizing one meaning of complexity for data. His computations verify that taking model uncertainty into account explicitly can lead to better predictions. According to his words, “if a Coordinating Theory could be found, it would serve the same role in statistics as Newton's Laws did for physics or evolution does for biology."

A View From a Limited Perspective was a title of the talk of Laurie Davies (University of Duisburg–Essen). It touched on various topics—the location-scale problem, approximation of a data set, topologies, likelihood, stability, regularization, the analysis of variance, sparsity, nonparametric regression, smoothness, shape, the role of asymptotics, optimality, theorems and procedures—on which this Davies has some experience, and also his unique personal perspective.

Another two presentations were devoted to random effects, a theme where "old statistics" has still much to say. Peter McCullagh (University of Chicago) in Random Effects and Estimating Equations, addressed the consequences for random-effects models of the sampling scheme or recruitment strategy used to enroll units. This issue does not usually arise in typical agricultural or horticultural field trials or laboratory experiments, but it does arise in clinical trials and in social-science areas such as marketing studies. McCullagh suggested a point-process model as a way to generate units in an automatic random manner, thereby avoiding the notion of units altogether.

Debashis Paul (Davis, jointly with Jie Peng and Prabir Burman) in Statistical Modeling Through Nonlinear Mixed Effects Dynamics focused on a class of models where the observations are longitudinal data measured for several subjects, and the trajectories are modeled as random and noisy realizations of a smooth process described by a first order nonlinear differential equation. Paul proposed a procedure based on numerically solving an ODE for the associated initial value problem to estimate the underlying dynamics from the observed data, studied the corresponding model selection problem, proposed an inferential framework for this class of problems, and provided illustrations of the method with simulated and real data examples.

Wolfgang Polonik (Davis, jointly with David Mason) in Asymptotic Normality of Plug-In Level Set Estimates spoke about level sets, regions where a target function f exceeds a given threshold value c. Such sets play a vital role in various fields of applications, such as anomaly detection, astronomical sky surveys, flow cytometry, and image segmentation. Other statistical applications of level sets include classification and visualization of multivariate densities. Algorithms have been devised for fast computation of level set estimates in large dimension, and consistency as well as optimal and 'fast' rates of convergence have been derived. While these results are interesting from a theoretical and computational point of view, they are not too helpful for statistical inference. Polonik addressed the problem of inference for level sets by focusing on a plug-in estimator (based on a kernel density estimator) of a density level set. As a distance measure, he considered the set-theoretic difference between the estimate and the target set, presented conditions under which such plug-in level set estimates are asymptotically normal, and discussed potential applications of such results to binary classification.

David Mason (University of Delaware, jointly with Julia Dony, Uwe Einmahl and Jan Swanepoel) discussed in Recent Results on Uniform in Bandwidth Consistency of Kernel-Type Function Estimators a general method based on empirical process techniques to prove uniform in bandwidth consistency of kernel-type function estimators. Examples include the kernel density and distribution function estimators, the Nadaraya-Watson regression function estimator, the conditional empirical process and conditional U-statistic. The results are useful to establish uniform consistency of data-driven bandwidth kernel-type function estimators.

Aurore Delaigle (University of Bristol) spoke about Design-Adaptive Local Polynomial Estimator for the Errors-in-Variables Problem, focusing on local polynomial estimators, a very popular techniques of nonparametric regression estimation that received great attention in the literature. Their simplest version, the local constant estimator, can be easily extended to the errors-in-variables context by exploiting its similarity with the deconvolution kernel density estimator. The generalization of the higher order versions of the estimator, however, is not straightforward and has remained an open problem for the last 15 years. In her talk, Delaigle showed how to construct local polynomial estimators of any order in the errors-in-variables context, discussed their asymptotic properties and illustrated their finite sample performance on numerical data examples.

Finally, Vladimir Vovk (Royal Holloway) in Predictive Regression: In Defence of the Old Statistics addressed the following problem: given the past data $(x_1, y_1), ..., (x_N, y_N)$ and a new vector x_{N+1} of explanatory variables, predict the new response variable y_{N+1}. The classical, and much criticized, assumption is the Gauss linear model: nothing is assumed about the explanatory vectors x_n, and the response y_n is modelled as a linear function of x_n plus an IID Gaussian noise. In the theoretical "new statistics" the Gauss linear model is usually replaced by the IID model: the observations (x_n, y_n) are independent and identically distributed. This greatly weakens the restriction on the distribution of responses given the explanatory vectors, but imposes restrictions on the distribution of explanatory vectors: for example, now they cannot be chosen by a free agent, and the prediction procedure cannot be used for control. The two models are not comparable: either has important advantages over the other. Vovk stated and discussed two relatively recent results. 1. Under the Gauss linear model and when used in the on-line mode, the classical prediction intervals based on Student's t-distribution fail to contain y_n independently for different n. Therefore, the chosen significance level translates into the frequency of error. 2. There exists an algorithm, Ridge Regression Confidence Machine, that satisfies analogous properties under the iid model: the probability of error is equal to the chosen significance level, and errors are made independently for different observations.

2.6 Object-oriented Data Analysis

Another recurring theme of the workshop was delineated by the one of the opening talks. Steve Marron (University of North Carolina) spoke about High Dimension Low Sample Size Mathematical Statistics by Steve Marron, the rapidly proliferating, but less mathematically analyzed theme—perhaps because the usual asymptotics are no longer relevant. He deems that a more appropriate HDLSS asymptotic theory, based on fixed sample size, with increasing dimension, is perhaps surprisingly relevant and useful. Results so far fall into two classes. The first is the discovery that, modulo rotation, random HDLSS data have a rigid deterministic structure, which reveals a number of useful statistical insights. The second is a class of results studying commonly used estimators, such as principal component direction vectors, are either consistent or strongly inconsistent (i.e. the angle between to the direction being estimated tends to 90 degrees), depending on the strength of the signal in the data.

Jim Ramsay (McGill), in Parameter Cascading for High Dimensional Models, spoke about high dimensional models that often involve three or more classes of parameters. Nuisance parameters are required to fit the data, are large in number, their number tends to depend on how much data is available, often define localized effects on the fit, and their values are seldom of direct interest. Structural parameters are the conventional kind; a small fixed number and their values are of interpretive importance. Above these are the complexity parameters that define the overall complexity of the solution. Ramsay defined a general framework for parameter estimation that synthesizes a variety of common approaches and brings some important new advantages. The parameter cascade approach involves defining nuisance parameters as functions of structural parameters, and in turn defines structural parameters as functions of complexity parameters. The method is much easier to program and tends to be much more computationally stable than classic marginalization approaches, and is an attractive alternative to MCMC.

Christopher Genovese (Carnegie-Mellon, jointly with Marco Pacifico Perone, Isabella Verdinelli, and Larry Wasserman), spoke about Finding Filaments, one-dimensional curves embedded in a point process or random field. He considered the problem of reliably detecting and estimating filaments, the problem arising in a wide range of applications. Statistical techniques exist for for finding one (or a few) filaments, but these methods do not handle noisy data sets with many filaments. Other methods can be found in the astronomy literature, but they do not have rigorous statistical guarantees. Genovese discussed two methods and their underlying theory. The first method locates filaments by finding regions where certain paths constructed from the data are highly concentrated. Concentration here refers to a formal density of paths. Genovese defined this density and constructed a consistent estimator of it. The second method combines nonparametric regression with vector quantization to partition the space into a collection of smooth curves. Genovese illustrated the procedure and assessed its theoretical performance. The two methods come from different theoretical directions and give different insights into the problem.

Martin Bilodeau (Université de Montréal) spoke about Discovering Dependencies in Multivariate Data and in Stationary Sequences. The Möbius transformation of probability cells in a multi-way contingency table is used to partition the Pearson chi-square test of mutual independence into A-dependence statistics. Bilodeau proposed a similar partition for a universal and consistent test of serial independence in a stationary sequence of a categorical variable, which can be adapted whether using estimated or theoretical marginal probabilities. With the aim of detecting a dependence of high order in a long sequence, A-dependence terms of the partition measuring increasing lagged dependences can be combined in a Box-Pierce type test of serial independence. Bilodeau presented a real data analysis of a nucleotides sequence using the Box-Pierce type test. A non parametric test of the mutual independence between many numerical random vectors is also proposed. This test is based on a characterization of mutual independence defined from probabilities of half-spaces in a combinatorial formula of Möbius. The critical values of the proposed test are computed with the bootstrap which is shown to be consistent. Another similar test, with the same asymptotic properties, for the serial independence of a multivariate stationary sequence is also proposed.

The workshop was concluded by the talk The Joy of Text by Andrey Feuerverger (Toronto), who introduced some new classes of statistical problems associated with computationally intensive statistical analysis of textual documents intended for a variety of purposes. These include the dating (or calendaring) of undated textual documents by comparing their contents with those of dated documents, and the sorting of a collection of documents into chronological order by means of comparing word sequences in the documents.

3 Scientific Progress Made

As mentioned at the beginning, the lessons learned at the meeting were partly expected, partly surprising. Several of the talks (Pötscher, McCullagh, Vovk, among others) underlined the continuity, the potential of the established statistical thinking if well understood and appropriately used. For instance, Benedikt Pötscher in his talk characterized himself as “a messenger from the past”; nevertheless, the impact of his talk concerned very recent themes. Summarizing this point, “old statistics” is not passé, only some of its too specialized disputes; and “new statistics” does not necessarily means “right” or “better”.

As expected, several areas of recent research confirmed themselves as particularly active and fertile directions: random matrices, sparse representation, geometric theory, regularization, algorithms. They may be in different stages of development and evolution; assessing those may be somewhat delicate and it is probably best to refer to the list of contributions for an ultimate prespective. Nevertheless, what may be said that each of these offers a promise of substantial research contributions in the years to come.

Of course, the progress in statistical sciences is by no means expected only in the just named areas. There are vast areas of very active research that were inevitably omitted at the meeting at this size; glimpses to other of them have been given by several talks. If one has to summarize here, the common prevailing focus would be less on rounding the angles and complementing the existing methodologies— not on “filling the gaps in the literature”, but rather on creating them; the economic law of diminishing returns seems to be fully applicable also here, and first insights and often approximate solutions are typically more appreciated as later improvements and refinements.

Another recurring theme is an emphasis on complexity, both of the data and methods of their analysis, and the subsequent management of this complexity by conceptual and algorithmic means. While the past methodologies amounted to fitting of and inferences about several numbers, slowly moving to lines and curves, the present practice of “object oriented data analysis” goes far beyond the “classical nonparametric perspectives”: the fitted concepts are collection of curves, algebraic structures like graphs and trees, and parameter cascades.

An important outcome of the meeting was the renewed recognition of the vital need of theoretical reflection. As mentioned already in the proposal, the rapidly growing needs of the statistical sciences provide raw material for future core research in statistics and motivates the development of trustworthy, user-friendly statistical methodology. However, as Sara van de Geer pertinently remarked, there may be way too many data analytic proposals out there; one desperately needs some insights. In this context, it is worth to repeat that statistics indeed fluctuates between import and export mode: importing raw data-analytic ideas inspired by the technology and problems of the moment and exporting refined data-analytic procedures, whose characteristics are understood theoretically and experimentally, to the community of quantitative scientists. The phrase “core of statistics” refers precisely to the intellectual basis for the export mode.

Finally, a positive feeling which was also to some extent felt in the community is that of the renewed awareness of common roots and inclinations. It turns out that the schism indicated at the beginning of this decade by [2] was perhaps slightly overstated, driven perhaps rather by a desire to be polemic with certain surviving tendencies at that time; it seems that the atmosphere in the data-analytic community at the end of this decade is much more about collaboration and unification; the divisive aspects seem to be less pronounced.

4 Outcome of the Meeting

Writing at this point that "the workshop has been a success" would be hardly a surprising twist in the context of BIRS final reports. While the feelings of the participants are probably on the high side, influenced also by the excellent weather we were lucky to enjoy, the full appreciation of the impact of the the workshop, its objectives and contents will be clear only with some time.

Nonetheless, there is one important objective that might have already been achieved, as stressed in the Tuesday evening round-table discussion by one of the organizers, Iain Johnstone. He recalled a recent workshop funded by NSF's Division of Mathematical Sciences at which speakers representing a variety of domains of statistical application emphasised the role of statistical theory in identifying, articulating and developing intersections in concepts and methods across the many areas in which statistical thinking is used. In this particular context, it is clear that the fact that a BIRS workshop like this—*planned and approved well before the NSF workshop was conceived*—not only happens, but convincingly demonstrates that cultivating statistical theory is a vital necessity for the healthy life of the statistical discipline, and that it is theory that often propels the applications—this fact alone may mean that the objectives of the meeting were met.

The workshop was attended by 41 confirmed (and present) participants, whose list is attached as appendix to this report. About 8–10 were in the "early pre-tenured" stage of their career (graduate students, postdocs, fresh assistant professors). We would like to extend our thanks to BIRS and its scientific director, Nassif Ghoussoub, for the opportunity to organize the workshop, and to all BIRS staff, especially to Alitha D'Ottavio and Brenda Williams, for help.

References

[1] R. Beran, Discussion of "Approximating Data" *Journal of the Korean Statistical Society* **37** (2008), 217–219.

[2] L. Breiman, Statistical modeling: The two cultures (with comments and a rejoinder by the author), *Statistical Science*, **16** (2001), 199-231.

[3] B. G. Lindsay, J. Kettenring, D. O. Siegmund, A report on the future of statistics, *Statistical Science*, **19** (2004), 387-413.

Appendix: List of participants

Belkin, Mikhail (Ohio State University)
Beran, Rudolf (University of California, Davis)
Bilodeau, Martin (Université de Montréal)
Brown, Lawrence (Larry) (University of Pennsylvania)
Bunea, Florentina (Florida State University)
Chen, Gemai (University of Calgary)
Chenouri, Shojaeddin (University of Waterloo)
Clarke, Bertrand (University of British Columbia)
Davies, Laurie (University of Duisburg-Essen)
Delaigle, Aurore (University of Bristol)
Doksum, Kjell (University of Wisconsin)
El Karoui, Noureddine (University of California Berkeley)
Farahmand, Amir massoud (University of Alberta)
Feuerverger, Andrey (University of Toronto)
Fukumizu, Kenji (Institute of Statistical Mathematics)
Genovese, Christopher (Carnegie Mellon University)
Hallin, Marc (Universite Libre de Bruxelles)
Hlubinka, Daniel (Charles University)
Hooker, Giles (Cornell University)
Johnstone, Iain (Stanford University)

Jordan, Michael (University of California Berkeley)
Kim, Peter (University of Guelph)
Koenker, Roger (University of Illinois at Urbana-Champaign)
Kovac, Arne (University of Bristol)
Leeb, Hannes (Yale University)
Maathuis, Marloes (Eidgenössische Technische Hochschule Zürich)
Marron, J. S. (Steve) (University of North Carolina Chapel Hill)
Mason, David M. (University of Delaware)
McCullagh, Peter (University of Chicago)
Mizera, Ivan (University of Alberta)
Owen, Art B. (Stanford University)
Paul, Debashis (University of California Davis)
Poetscher, Benedikt M. (University of Vienna)
Polonik, Wolfgang (University of California Davis)
Rajaratnam, Bala (Stanford University)
Ramsay, Jim (McGill University)
Rosset, Saharon (Tel Aviv University)
van de Geer, Sara (Eidgenössische Technische Hochschule Zürich)
Vovk, Vladimir (Royal Holloway, University of London)
Wegkamp, Marten (Florida State University)
Zlatev, Boyko (University of Alberta)