STATISTICS AT THE FRONTIERS OF SCIENCE

Gemai Chen (The University of Calgary),
David R. Brillinger (University of California at Berkley),
Jianqing Fan (Princeton University),
Jun Liu (Harvard University),
James O. Ramsay (McGill University),
Keith J. Worsley (McGill University)

June 24–June 29, 2006

1 Overview of the Field

Statistics may be broadly defined as the theory and methods of collecting, organizing and interpreting data for solving real world problems. The problems may be formulated as testing of a hypothesis, choosing between alternative hypotheses, estimating an unknown quantity, predicting a future event, and in general, making a decision under uncertainty with minimum risk or loss.

Historically, the ideas and methods of statistics developed gradually as society grew interested in collecting and using data for a variety of applications. The earliest origins of statistics lie in the desire of rulers to count the number of inhabitants or measure the value of the taxable land in their domains. As the physical sciences developed in the 17th and 18th centuries, the importance of careful measurements of weights, distances, and other physical quantities grew. Astronomers and surveyors striving for exactness had to deal with variation in their measurements. Many measurements should be better than a single measurement, even though they vary among themselves. How can we best combine many varying observations? Statistical methods that are still important were invented in order to analyze scientific measurements.

By the 19th century, the agricultural, life, and behavioral sciences also began to rely on data to answer fundamental questions. How are the heights of parents and children related? Does a new variety of wheat produce higher yields than the old, and under what conditions of rainfall and fertilizer? Can a person's mental ability and behavior be measured just as we measure height and reaction time? Effective methods for dealing with such questions developed slowly and with much debate.

As methods for producing and understanding data grew in number and sophistication, the new discipline of statistics took shape in the early part of the 20th century. Ideas and techniques that originated in the collection of government data, in the study of astronomical or biological measurements, and in the attempt to understand heredity or intelligence came together to form a unified "science of data". As huge computing power has become more and more accessible in the past two decades or so, the complexity of our society has increased dramatically, the amount of relevant information has exploded, and statistics has become more and more essential in reaching a scientific decision.

Most importantly, what we can see from the above overview is that it is the application that gives new life to statistics, and at the same time, statisticians have been helping scientists in solving current problems and formulating new theories that will lead to advancement of knowledge.

2 The Objectives of the Workshop

The workshop has invited active statistical researchers involved in various areas of scientific research to meet and exchange ideas. Our objectives are: (1) to bring the inspiring excitement of frontier scientific research to statistical community to strengthen the current statistical research, and (2) to create opportunities for new and/or deeper collaborative research.

Five group leaders have participated in the invitation, organization and running of the workshop. They are

- Professor David R. Brillinger of University of California at Berkley responsible for **Time Series and Stochastic Processes** (Due to other commitment, Professor Brillinger was not able to attend the workshop after organizing his group. Professor Bruce Smith took over the responsibility during the workshop.)
- Professor Jianqing Fan of Princeton University responsible for **Financial and Risk Analysis**
- Professor Jun Liu of Harvard University responsible for **Bio-Medical Research**
- Professor James O. Ramsay of McGill University responsible for **Functional Data Analysis**
- Professor Keith J. Worsley McGill University responsible for **Random Field and Image Analysis**

The above five areas represent a sample of the areas where statisticians are working closely with frontier scientific researchers to attack some of the most challenging problems today. Taking a function or a spectrum as input instead of a single numerical datum, functional data analysis methods have evolved from science and are now helping scientists to research from a broader and more realistic perspective. Advances in geometry and random fields driven by our desire to understand human being better have made it possible to quantify some topological changes, such as those in brain shape. The ups and downs of the financial market have definitely stimulated the statistical research to come up new models, such as the two proposed by the two 2003 Nobel Prize winners in economics. The great efforts made to understand the origin of life and develop new and more effective drugs to battle diseases have changed the way statistical research is usually done, offered a whole range of challenging problems, and made statistics an integral part of the biological and medical research. The innovations and comebacks of new and old techniques in time series and stochastic processes have merged statistics into many scientific endeavors that are changing the face of science day after day.

The various annual statistical meetings do include the above groups in the programs, but the busy schedules usually prevent the different groups to sit together and share ideas. This workshop has provided such an opportunity. More importantly, the seemingly unrelated groups actually have much in common from a methodological point of view, and good opportunities have been taken for transferring, borrowing and strengthening already developed methodologies, and for creating new research directions.

3 Presentation Highlights

In the following, we will highlight the various presentations given during the workshop according to the 5 different topics.

3.1 Financial and Risk Analysis

The presentation of Professor Per Mykland (University of Chicago, mykland@galton.uchicago.edu) titled “A Gaussian Calculus for Inference from High Frequency Data” is aimed at providing a rigorous theory for some of the existing methodology and possible new tools for financial analysis. In the econometric literature of high frequency data, it is often assumed that one can carry out inference conditionally on the underlying volatility processes. In other words, conditionally Gaussian systems are considered. This is often referred to as the assumption of “no leverage effect”. This is often a reasonable thing to do, as general estimators and results can often be conjectured from considering the conditionally Gaussian case. This presentation is to try to give some more structure to the things one can do with the Gaussian assumption. It is argued that there is a whole

treasure chest of tools that can be brought to bear on high frequency data problems in this case. In particular, approximations involving locally constant volatility processes are considered, and a general theory for this approximation is developed. As applications of the theory, an improved estimator of quarticity, an ANOVA for processes with multiple regressors, and an estimator for error bars on the Hayashi-Yoshida estimator of quadratic covariation are proposed.

Professor Yazhen Wang (University of Connecticut, yzwang@stat.uconn.edu) delivers a talk on "Heterogeneous Autoregressive Realized Volatility Model". Volatilities of asset returns are pivotal for many issues in financial economics such as asset pricing, portfolio allocation and risk management. The availability of high frequency intraday data may allow people to estimate volatility more accurately. In practice, realized volatility is often used to estimate integrated volatility. To obtain better volatility estimation and forecast, some autoregressive structure of realized volatility is proposed in the literature. The use of a heterogeneous autoregressive model for realized volatility is explored and a nonparametric multiscale statistical technique is developed to construct noise resistant realized volatility for estimating integrated volatility.

"Statistical Approaches to Option Pricing and Portfolio Management" delivered by Professor Jianqing Fan (Princeton University, jqfan@princeton.edu) addresses the fundamental issues of finance. The existing financial mathematical models provide useful tools for option pricing. These physical models give us a good first order approximation to the underlying dynamics in the financial market. However, their power in option pricing can be significantly enhanced when they are combined with statistical approaches, which empirically learn and correct pricing errors through estimating the state price densities. Two new semiparametric techniques are proposed for estimating state price densities and pricing financial derivatives. Empirical studies based on the options of SP500 index over 100,000 tests show that the two new semiparametric techniques outperform the ad hoc Black-Scholes method and significantly so when the latter method has large pricing errors.

A related issue is to find a good estimation of the high-dimensional covariance matrix for portfolio allocation and risk management. Motivated by the Capital Asset Pricing Model, a factor model is proposed to reduce the dimensionality and to estimate the covariance matrix. The performance of the new estimate is compared with the sample covariance matrix. Situations under which the factor approach can gain substantially in performance and the cases where the gains are only marginal are demonstrated and identified. Furthermore, the impacts of the covariance matrix estimation on portfolio allocation and risk management are studied. The theoretical results are convincingly supported by a thorough simulation study.

An interesting feature of the talk is that traditionally, people seek to have the best, the most, or the perfect. But in applications involving complicated phenomena such a finance, it is more beneficial to seek a better, a useful, or an improved result.

Professor Cheng-Der Fuh (Academia Sinica, stcheng@stat.sinica.edu.tw) gives an interesting and deep talk on "Efficient Likelihood Estimation in State Space Models". Likelihood principle is the most used principle is statistical theory and applications. Motivated by studying asymptotic properties of the maximum likelihood estimator (MLE) in stochastic volatility (SV) models, likelihood estimation in state space models is investigated. It is first proved that under some regularity conditions, there is a consistent sequence of roots of the likelihood equation that is asymptotically normal with the inverse of the Fisher information as its variance. With an extra assumption that the likelihood equation has a unique root for each n, then there is a consistent sequence of estimators of the unknown parameters. If, in addition, the supremum of the log likelihood function is integrable, the MLE exists and is strong consistent. Edgeworth expansion of the approximate solution of likelihood equation is also established. Several examples, including Markov switching models, ARMA models, (G)ARCH models and stochastic volatility (SV) models, are given for illustrations.

An interesting feature of the talk is that the essence of the well known maximum likelihood method is given a much simplified discussion and one can now see what needs to be done to use this method regardless of the specific details.

3.2 Time Series and Stochastic Processes

"Structural Break Detection in Time Series Models" are addressed by Professor Richard A. Davis (Colorado State University, rdavis@stat.colostate.edu). Much of the recent interest in time series modeling has focused on data from financial markets, from communications channels, from speech recognition and from engineering applications, where the need for non-Gaussian, non-linear, and nonstationary models is clear. With faster

computation and new estimation algorithms, it is now possible to make significant in-roads on modeling more complex phenomena. In this talk, Professor Davis develops estimation procedures for a class of models that can be used for analyzing a wide range of time series data that exhibit structural breaks. The novelty of the approach taken here is to combine the use of genetic algorithms with the principle of minimum description length (MDL), an idea developed by Rissanan in the 1980s, to find "optimal" models over a potentially large class of models.

This methodology is demonstrated in a number of applications including piece-wise AR models, segmented GARCH models, slowly varying AR models, linear models with dynamic structures and state space models. In addition to fitting piece-wise autoregressive models, which works well even for local stationary models that are smooth, extensions to piece-wise nonlinear models including stochastic volatility and GARCH models are also considered.

Professor R.H. Shumway (University of California at Davis, rhshumway@ucdavis.edu) talks about "Mixed Signal Processing for Regional and Teleseismic Arrays". Successful monitoring of a proposed Comprehensive Nuclear Test-Ban Treaty (CTBT) ultimately rests on interpretations of time series that are produced on seismic and infrasound arrays as well as on auxiliary information from other sources such as satellites and radionuclide sampling. Underground events such as earthquakes and explosions generate plane waves propagating across arrays of seismometers and proper use of this information is critical to the successful detection, location and identification of the source phenomenon.

When simultaneous events occur or when propagating noises are present at an array, mistakes can be made in locating an event as well as in reading the magnitude-related variables that are critical for discriminating between classes of events. The performance of conventional high-resolution estimators such as MUSIC for two typical mixtures of signals is examined and an alternate approach using a combination of nonlinear stepwise regression and model selection techniques is developed. The new method yields the correct number of signals on two typical mixtures and allows deconvolution of the component signals.

Peter Buhlmann (ETH Zurich, buhlmann@stat.math.ethz.ch) takes on the challenging "DNA Splice Site Detection with Group-penalty Methods for Categorical Predictors". DNA splice sites detection has been pursued, among other approaches, by non-Markovian time series models. As an alternative, Professor Buhlmann uses logistic regression with (short) DNA sequence as categorical predictors. When including higher-order interaction terms, such models become very high-dimensional (e.g. 1'000 - 16'000 predictors). He proposes to use the group-penalty and new modifications thereof for hierarchical model fitting and presents efficient algorithms which are particularly suited for high-dimensional problems. He shows that the proposed methods are statistically consistent for sparse but high-dimensional problems where the number of predictor variables may be much larger than sample size. Despite the generality of the new approach, it performs surprisingly well for the DNA splice site detection problems which he has analyzed.

Professor Ian McLeod (University of Western Ontario, aim@stats.uwo.ca) addresses the workshop with "My Current Research in Time Series". He starts with his (1) recent work on statistical algorithms for time series including subset autoregressive modeling, faster ARMA maximum likelihood estimation and automatic Brillinger monotonic trend test. He then moves to (2) improved portmanteau diagnostic checks for univariate and vector ARMA time series, followed by (3) applications of time series in bio-informatics. Finally he discusses a new diagnostic check for lack of statistical independence which is applicable to a wide variety of statistical models including regression and generalized linear models.

This time series and stochastic processes group ends with the talk by Wai Keung Li (The University of Hong Kong, hrntlwk@hku.hk) on "Least Absolute Deviation Estimation for Fractionally Integrated Autoregressive Moving Average Time Series Models with Conditional Heteroscedasticity". In order to model time series exhibiting the features of long memory, conditional heteroscedasticity and heavy tails, a least absolute deviation approach is considered to estimate fractionally autoregressive integrated moving average models with conditional heteroscedasticity. The time series generated by this model is short memory or long memory, stationary or nonstationary, depending on whether the fractional differencing parameter $d \in (-1/2, 0)$ or $(0, \infty)$, $d \in (-1/2, 1/2)$ or $(1/2, \infty)$ respectively. Using a unified approach, the asymptotic properties of the least absolute deviation estimation are established. The large sample distribution of residual autocorrelations and absolute residual autocorrelations is also derived and these results lead to two useful diagnostic tools for checking the adequacy of the ftted models. Some Monte Carlo experiments were conducted to examine the performance of the theoretical results in finite sample cases. As an illustration, the process of modeling the absolute return of the daily closing Dow Jones Industrial Average Index (1995-2004) is also reported.

3.3 Functional Data Analysis

In an overview address titled "Functional Data Analysis: Where it's been and where it might be going?", Professor Jim Ramsay (McGill University, ramsay@psych.mcgill.ca) reviews the history and current trends in functional data analysis: where the field tends to fit in with respect to other methods looking a distributed data, what the more important accomplishments have been over the last decade, current work on modeling dynamic systems, and what the future might hold.

For example, over the past two years Professor Ramsay's research has been aimed at taking functional data analysis into the world of dynamic systems modeling. This means developing the capacity to estimate the parameters defining a system of differential equations, either linear or nonlinear, from noisy data, often taken from only a subset of the variables in the system. This has worked out very well, and involves using the DIFE to define a roughness penalty, and then using a two- or three-stage estimation procedure in which:

- coefficients of basis functions expansions for variables are defined as functions of parameters defining the system, and are estimated conditional on system parameters by optimizing a fit measure plus roughness penalty
- defining system parameters as functions of smoothing or other model complexity parameters, and each time these latter are changed, optimizing a measure of fit without regularization and, finally
- optimizing an estimate of mean square error such as GCV with respect to smoothing parameters.

This three-stage process involves what Professor Ramsay has come to call a "parameter cascade" in which parameters are segmented into levels, with the lowest level being "local parameters", the next level "global parameters", and the highest level "complexity parameters." It was this perspective that finally really made our many successful estimates of DIFE systems possible.

The parameter cascade idea has immediate application to many other situations. In particular, it deals with the famous Neyman-Scott problem very nicely, applies neatly to linear and nonlinear mixed effects or multi-level models, works well with psychometric problems involving estimating examinee ability and item characteristics, and so on. In effect, it is a Bayesian like framework that leads to much faster and more direct methods than MCMC that also have easily computable interval estimates. Unlike MCMC, it is easy to deploy to users.

In an coordinated talk titled "Functional Data Analysis: Tools and issues", Professor Hans-Georg Mueller (University of California at Davis, Mueller@wald.ucdavis.edu), Professor Daniel Gervini (University of Wisconsin at Milwaukee, gervini@uwm.edu), Professor Fang Yao (Colorado State University, ffyao@stat.colostate.edu) and Professor Gareth James (University of Southern California, gareth@usc.edu) explore a range of topics in functional data analysis.

Professor Mueller opens the talk by discussing the characteristics of functional data: High-dimensional (infinite-dimensional) with a topology characterized by order, neighborhood and smoothness, and various warping methods—curve alignment or registration. He also addresses the issues of functional convex calculus and weak convergence.

Professor Gervini picks up functional principal component analysis which is a generalization of the usual principal component analysis for multivariate data. From Karhunen-Lóeve expansion, to asymptotic properties, to functional longitudinal sparse data analysis, and to future studies (spacial and image analysis, non-Gaussian longitudinal data, high-dim matrices) Professor Gervini gives a complete summary of functional principal component analysis.

Professor Fang Yao discusses various functional regression models including multivariate regression, functional linear model, functional response model, and generalized functional linear model.

Professor James reports on the challenges when developing and applying functional data analysis methodologies (increasingly complex functional regression models, inference and asymptotics in new settings, survival analysis (joint modeling), ecology, financial time series, and gene expression time courses).

A feature of the above talks in this group is that it is firmly believed that the field will be driven forward by new applications.

3.4 Random Field and Image Analysis

Among the 5 topics of the workshop, this topic seems to be the most abstract, but the applications presented are just astonishingly concrete.

Professor Robert J. Adler (Technion - Israel Institute of Technology, robert@ieadler.technion.ac.il) gives the first talk on "Rice and Geometry". The classic Rice formula for the expected number of upcrossings of a smooth stationary Gaussian process on the real line is one of the oldest and most important results in the theory of smooth Gaussian processes. It has a multitude of applications, and has been generalized over the years to non-stationary and non-Gaussian processes, both over the real and over more complex parameter spaces, and to vector valued rather than real valued processes.

Over the last few years, a new Rice"super formula" has been developed, which incorporates effectively all the (constant variance) special cases known until now. More interesting, however, is that this new formula shows that all of the related formulae have a deep geometric interpretation, giving a version of the Kinematic Fundamental Formula of Integral and Differential Geometry for Gauss space.

Later on Professor Adler outlines a proof of the new formula and says: "There are two choices. One is to use Riemann metric and geometry; the other is to rediscover Riemann metric and geometry."

Professor Keith Worsley (McGill University, keith@math.mcgill.ca) speaks on behalf of Dr. Jonathan Taylor (Stanford University, jonathan.taylor@stanford.edu) on "Deformation Based Morphometry, Roy's Maximum Root and Recent Advances in Random Fields". He starts with a study of anatomical differences between controls and patients who have suffered non-missile trauma. He then employs a multivariate linear model at each location in space, using Hotelling's T^2 to detect differences between cases and controls. If covariates are further included in the model, Roy's maximum root is a natural generalization of Hotelling's T^2. This leads to the Roy's maximum root random field, which includes many special types of random fields: Hotelling's T^2, T, and F, so, in effect the Roy's maximum root random field "unifies" many different random fields. The geometric interpretations of this "unified" random field theory is explored.

Professor Jiayang Sun (Case Western Reserve University, jiayang@sun.STAT.cwru.edu) continues to talk about "Three Statistical Imaging Problems". First, she describes two neuron-imaging problems ((i) ctivities in spinal cord: Fos expression, involving cats and measured with spatial counts of neurons activated by stimulated nerves; (ii) changes in brain activities: force, EMG (3), EEG (64) involving human and measured with structured time series (fat and short)) that challenge the status quo, namely, statistical models assumed for typical neuronal data. These problems offer opportunities for new modeling and statistical inference, including multiple comparisons arising from a negative binomial random field, which is generally applicable for analyzing data from over-dispersed Poisson regression models. The third imaging problem that led to her research on identifying the pixels that most likely correspond to the false discoveries from a FDR procedure is also touched.

In the talk titled "Granger Causality on Spatial Manifolds: Applications to Neuroimaging", Professor Pedro A. Valds-Sosa (Cuban Neuroscience Center, Ciudad Habana, Cuba, peter@cneuro.edu.cu) combines economics with neuroscience. The (discrete time) vector Multivariate Autoregressive (MAR) model is generalized as a stochastic process defined over a continuous spatial manifold. The underlying motivation is the study of brain connectivity via the application of Granger Causality measures to functional Neuroimages. Discretization of the spatial MAR (sMAR) leads to a densely sampled MAR for which the number of time series p is much larger than the length of the time series N. In this situation usual time series models work badly or fail. Previous approaches, involve the reduction of the dimensionality of the MAR, either by the selection of arbitrary regions of interest or by latent variable analysis. An example of the latter is given using a multi-linear reduction of the multichannel EEG spectrum into atoms with spatial, temporal and frequency signatures. Influence measures are applied to the temporal signatures giving an interpretation of the interaction between brain rhythms. However the approach introduced here is that of extending the usual influence measures for Granger Causality to sMAR by defining "influence fields", that is the set of influence measures from one site (voxel) to the whole manifold. Estimation is made possible by imposing Bayesian priors for sparsity, smoothness, or both on the influence fields. In fact, a prior is introduced that generalizes most common priors studied to date in the literature for variable selection and penalization in regression. This prior is specified by defining penalties paired with a priori covariance matrices. Simple pairs of penalties/covariances include as particular cases the LASSO, Data Fusion and Ridge Regression. Double pairs encompass the recently introduced Elastic Net and Fussed Lasso. Quadruples of penalty/covariance combinations are also

possible and used for the first time. Estimation is carried out via the MM algorithm, a new technique that generalized the EM algorithm and allows efficient estimation even for massive time series dimensionalities. The proposed technique performs adequately for a simulated "small world" cortical network with linear dynamics, validating the use of the more complex penalties. Application of this model to fMRI data validate previous conceptual models for the brain circuits involved in the generation of the EEG alpha rhythm.

The last talk in this group given by Professor Emery N. Brown (Massachusetts Institute of Technology, brown@neurostat.mgh.harvard.edu) on "Large Scale Kalman Filtering Solutions to the Electrophysiological Source Localization Problem—An MEG Case Study" extends the use of the famous Kalman filter to a new application. Computational solutions to the high-dimensional Kalman Filtering problem are described in the setting of the MEG inverse problem. The overall objective is to localize and estimate dynamic brain activity from observed extraneous magnetic fields recorded at an array of sensor positions on the scalp and to do so in a manner that takes advantage of the true underlying statistical continuity in the current sources. To this end, one can use inverse mapping procedures that combine models of current dipoles with dynamic state-space estimation algorithms. While these algorithms are eminently well-suited to this class of dynamic inverse problems, they possess computational limitations that need to be addressed either by approximation or through the use of high performance computational resources. A High Performance Computing (HPC) solution to the Kalman filter is found and its applicability to the Magnetoencephalography (MEG) inverse problem is demonstrated.

3.5 Bio-Medical Research

This is a very dynamic group attacking various bio and life science challenges.

Professor Jun Liu (Harvard University, jliu@stat.harvard.edu) delivers a talk on "A Bayesian Method for Detecting Disease-Related Genetic Interactions". In case-control association studies, it is of interest to detect multi-locus interactions called Epistasis. More specifically, given genotypes at multiple loci for both cases and controls, one would like to locate most likely positions where a disease-related mutation may have occurred. Various parametric and nonparametric methods are reviewed and it is noted that the existing methods are either of low power or computationally infeasible when facing a large number of markers. An alternative method using MCMC sampling techniques is developed which can efficiently detect interactions among thousands of markers. The issues of statistical significance and how to adjust multiple comparisons are discussed (much of these are conjectures, though).

"Using Genetic Linkage to Inform Positional Cloning", Professor Mary Sara McPeek (University of Chicago, mcpeek@galton.uchicago.edu) addresses an delicate issue in genomics. The first step in mapping a gene for a trait often involves using linkage analysis to identify a region on a chromosome where a gene of interest may lie. Linkage disequilibrium mapping may sometimes be used to further refine the region. At this point, one may be able to identify one or several genes within the region. Even if only a single gene lies in the region, it may contain a large number of polymorphic sites (base pairs of DNA that vary across individuals), and a question of interest is to determine which site or combination of sites influence the trait. Ultimately, only well-designed biological studies can establish that particular variation influences susceptibility. However, one can address the question of whether a particular set of polymorphisms can fully "explain", in the statistical sense, the observed linkage to the region. Suppose that many tightly-linked SNPs have been identified and genotyped in affected relatives in a region showing strong linkage with a binary trait. It is noted that if a particular set of SNPs contains all the sites in the region that influence the trait, then conditional on the genotypes at those SNPs, there should be no excess sharing in the region among affected individuals. This idea is used to develop a statistical test of the null hypothesis that a particular SNP or pair of SNPs can explain all the evidence for linkage and to develop a confidence set of individual SNPs and pairs of SNPs that has appropriate coverage probability of the causal SNP or pair of SNPs assuming that there are no more than 2 in that region. Arbitrary genetic model (including epistasis with other unlinked susceptibility loci) and linkage disequilibrium are allowed and appropriate methods to take into account the uncertainty in haplotype frequencies re developed. Discussions on approaches to the problem of adjusting for the selection of the region based on the linkage results in the same sample of individuals are offered.

Professor Hongyu Zhao (Yale University, hz27@email.med.yale.edu) entertains a challenging issue in microbiology to talk about "Protein Interaction Predictions Through Integrating High-throughput Data From Diverse Organisms". Predicting protein-protein interactions is critical for understanding cellular processes.

Because protein domains represent binding modules and are responsible for the interactions between proteins, several computational approaches have been proposed to predict protein interactions at the domain level. The fact that protein domains are likely evolutionarily conserved allows us to pool information from data across multiple organisms for the inference of domain-domain and protein-protein interactions. Professor Zhao presents his results on estimating domain-domain interaction probabilities through integrating large-scale protein interaction data from three organisms, yeast, worms, and fruit flies. The estimated domain-domain interaction probabilities can be then used to predict protein-protein interactions in a given organism. Based on a thorough comparison of sensitivity and specificity, and other analyses, the proposed approaches are shown to have better performance due to their ability to borrow information from multiple species. The estimated domain-domain interaction probabilities can also be informative in predicting protein-protein interaction in other organisms.

Recent advances in life science have allowed scientists to "follow" the movement of a molecule. Professor Samuel Kou (Harvard University, kou@stat.harvard.edu) introduces the audience to this fascinating topic with a talk titled "Stochastic Modeling and Inference in Nano-scale Biophysics". With the progress made in nanotechnology, scientists now can follow a biological process on an unprecedented single molecule scale. These advances also raise many challenging stochastic modeling and statistical inference problems. First, by zooming in on single molecules, recent nano-scale experiments reveal that some classical stochastic models derived from oversimplified assumptions are no longer valid. Second, the stochastic nature of the experimental data and the presence of latent processes much complicate the statistical inference. Professor Kou uses the modeling of subdiffusion phenomenon in enzymatic conformational fluctuation and the inference of DNA hairpin kinetics to illustrate the statistical and probabilistic challenges in single-molecule biophysics.

Professor Ying Nian Wu (University of California at Los Angeles, ywu@stat.ucla.edu) moves on to discuss an important technology in bio industry with a talk on "ChIP-chip: Data, Model, and Analysis". ChIP-chip (or ChIP-on-chip) is a technology for isolation and identification of genomic sites occupied by specific DNA binding proteins in living cells. ChIP-chip data can be obtained over the whole genome by tiling arrays, where a peak in the signal is generally observed at a protein binding site. Professor Wu presents a probability model for ChIP-chip data. Then he proposes a model-based computational method for locating and testing peaks for the purpose of identifying potential protein binding sites and presents a non-parametric method for identifying and representing peaks in multiple resolutions.

Cancer research is not new but is one of the most important and challenging bio-medical researches. Professor Volker Schmidt (UniversitŁt Ulm, volker.schmidt@uni-ulm.de) presents a talk on "Model-Based Analysis of Keratin Filament Networks in Scanning Electron Microscopy Images of Cancer Cells". The keratin filament network is an important part of the cytoskeleton in epithelial cells. It is involved in the regulation of shape and viscoelasticity of the cells. In-vitro studies indicated that geometrical network characteristics, such as filament cross-link density, determine the biophysical properties of the filament network.

Scanning electron microscopy images of filaments were processed by a skeletonisation algorithm based on morphological operators to obtain a graph structure which represents individual filaments as well as their connections. This method was applied to investigate the effects of the so-called transforming growth factor alpha (TGF-alpha) on the morphology of keratin networks in pancreatic cancer cells. By estimating geometrical network characteristics, like the length and orientation distributions of the keratin filaments, and by fitting random tessellation models, a significant alteration of keratin network morphology could be detected in response to TGF-alpha.

Professor Murray D. Burke (University of Calgary, burke@math.ucalgary.ca) gives the last talk of the workshop on "Semiparametric Regression Models with Staggered Entries and Progressive Multi-Stage Censoring". In his talk, Professor Burke studies a class of semiparametric regression models when subjects enter the study in a staggered fashion. A strong martingale approach is used to model the two-time parameter counting processes. It is shown that well-known univariate results such as weak convergence and martingale inequalities can be extended to these two-dimensional models. Strong martingale theory is also used to prove weak convergence of a general weighted goodness-of-fit process and its weighted bootstrap counterpart. If three progressive multi-stage censoring schemes are considered, where the experimenter purposely censors a given number of individuals under study at fixed time points, it is also possible to incorporate this censoring into the above models.

4 Scientific Progress Made

Although there are 5 groups of different researchers attending the workshop, both within the groups and among the groups, participants have had a wide range of (long) discussions.

The important role that statistics can play in financial and risk analysis is clearly demonstrated and enhanced. Finance theory based on (stochastic) partial differential equations can be used to set up the market and make it run. Statistical methodologies can validate/help adjust the financial theory, help us to understand financial volatility better, and help run the market more effectively.

To treat today's time series, new methods are clearly needed, especially in dealing with multivariate series involving both time and space. On the other hand, existing methods can be made more useful through combining them appropriately by identifying the structural breaks in the series.

Science nowadays offers more and more challenging problems and human beings have been able to take more and more complicated measurements. In just over a dozen years, functional data analysis has quickly become a powerful methodology for scientists to attack various old and new problems. This is well reviewed and illustrated in the workshop. If one wants to start a career in statistics, functional data analysis should be seriously considered.

Abstract and concrete are two extremes. When both are sophisticated, it is intimidating to try to understand them. However, when the two are coherently connected, it creates stunning beauty. The random field and image analysis group did just that. On the one side is the boundary crossing in Gaussian random space. On the other side is the change of brain tumors. Successful statistical procedures have been developed to quantify the change of brain tumors using the fundamental results from boundary crossing.

Life takes different forms at different levels. It therefore offers challenges of various kinds. Observations or measurements we can take all contain errors or uncertainties. With errors or uncertainties present, how to detect disease-related genetic interactions, how to use genetic linkage to inform positional cloning, how to predict protein interactions, how to model subdiffusion phenomenon in enzymatic conformational fluctuation, how to isolate and identify genomic sites occupied by specific DNA binding proteins in living cells, how to create a graph which represents individual filaments as well as their connections in cancer cells, and how to allow and handle censoring in medical research? All of these are directly related to whether we can make progress in understanding life. Statistical thinking and methods have been shown to be valuable for the above and other efforts.

Throughout the workshop, it has been demonstrated again and again and in one field after another that science offers new challenges to statisticians and statisticians can help make genuine and significant progress in science.

5 Outcome of the Meeting

There is no doubt that the workshop is a great success. The relaxed atmosphere and the plenty of time for discussion have allowed participants to share and exchange ideas, discuss issues of mutual interest, and team up to workshop on existing and new problems.

The discussion part has really stood out: in each of the 5 discussion periods for the 5 different groups, there are always participants from different groups to attend and the discussions are always interesting and entertaining.

Thanks to the wonderful and efficient management of BIRS, the workshop runs very smoothly. Here are some of the feedbacks:

"Thank you very much for all the work you did in organizing this retreat and conference. Professionally, it was the best workshop or conference I have ever attended. This was due to the relatively small number of participants, the beautiful location and the large number of social acitivities that you and the others organized. It was a wonderful experience."

"I just came back, and the first thing I want to do is to sincerely thank you and your colleagues for organizing such a nice workshop. I learned a lot, and I also thoroughly enjoyed Banff."

"Many thanks for organizing a fantastic event."

"Thanks so much for taking care of everything and running such an excellent conference."

“Thanks very much for organizing the workshop. I had a very good time and was able to make contact with some researchers previously admired only from afar.”

“I have returned home safe and sound. Many thanks again for the invitation and the organising. It was a great workshop and nice environment. I certainly have learned a lot.”

“Thank you for organizing the conference. It was a wonderful success. Thanks also go to your hospitality. I enjoy very much the wonderful Banff.”

“Thank you very much for having organized this great workshop: it has been very interesting (talks and discussions), it gave me the opportunity to talk “in depth” with many people and I also enjoyed the outdoors around Banff.”

“BIRS is better than Oberwolfach in environment, food and management.”