The Impact of Statistical Thinking in Economics and Life Sciences

Workshop in honour of Pietro Muliere

Milan, September 13-15, 2012

Schedule & Program

The Impact of Statistical Thinking in Economics and Life Sciences

Workshop in honour of Pietro Muliere

Room N03, piazza Sraffa 13

Schedule overview

	THURSDAY, Sept. 13	FRIDAY, Sept. 14	SATURDAY, Sept. 15
9:00-9:30	**Registration and welcome**		
9:30-9:45	**Opening**		
9:45-10:15	**Karl Mosler** Dual stochastic dominance and multidimensional inequality	**Marco Scarsini** On information distortion in online ratings	
10:15-10:45	**Eugenio Regazzini** Central limit theorem and asymtotics for solutions of kinetic equations	**Yosef Rinott** Cross sectional sampling, bias and dependence	**Eduardo Gutiérrez Peña** Proper and non-informative conjugate priors for exponential family models **(10:00-10:30)**
10:45-11:05	**Stefano Favaro** On the stick-breaking representation for Gibbs-type priors	**Chiara Gigliarano** Some results on stochastic comparisons of ROC curves	**Valen Johnson** Uniformly Most Powerful Bayesian Tests **(10:30-11:00)**
11:05-11:30	**Coffee break**	**Coffee break**	**Coffee break**
11:30-12:00	**Jon Wellner** Chernoff's distribution is log-cave… and more	**Pierluigi Conti** On the estimation of distribution functions in sampling finite populations	**Claudio Zoli** The Measurement of Multi-group Dissimilarity and Related Orders
12:00-12:30	**B.L.S.Prakasa Rao** Statistical inference for fractional diffusion processes	**Luca Tardella** Bayesian Poisson Mixtures for Frequencies of Frequencies	**Piercesare Secchi** Reinforced urns: a biased overview
12:30-13:00	**Peter Mueller** A Nonparametric Bayesian Model for Local Clustering	**Giovanni Parmigiani** Gene Set Analysis as a Tool for Cross-Platform Integration in Genomics	**Closing**
13:00-14:30	**Bocconi Buffet**		
14:30-15:00	**Michele Guindani** Bayesian Non parametric identication of Pre-versus-Post treatment biomarker effects on Progression free survival	**Nicholas Polson** Mixture Importance Splitting. Normalisation and Rare Events	
15:00-15:30	**Matteo Ruggiero** Bayesian nonparametric priors and Markov processes	**Dario Spano'** Canonical correlations for dependent random measures	
15:30-16:00	**Coffee break**	**Coffee break**	
16:00-16:20	**Sara Wade** More Informative Conditional Density Estimation with Multivariate Covariates	**Lorenzo Trippa** Bayesian Nonparametric Analysis of reversible Markov Chains	
16:20-16:40	**Pierpaolo De Blasi** Bayesian estimation of discrepancy with misspecified parametric models	**Antonietta Mira** Stochastic orderings: the sequel	
16:40-17:00	**Free contributions**	**Clelia di Serio** Simpson's Paradox in Survival Models	
17:00-17:30	**Free contributions**		
Evening		**Social Dinner**	

Thursday, September 13th

9:00-9:30 **Registration and Welcome**

9:30-9:45 **Opening**

9:45-10:15 **Dual stochastic dominance and multidimensional inequality**

Karl Mosler (University of Cologne, Germany)

Abstract: Pietro Muliere's by far most quoted work (Google Scholar shows 120 citations) is the short paper he wrote with Marco Scarsini, "A Note on Stochastic Dominance and Inequality Measures" (JET 49(1989), 314-323).There the two authors introduce a sequence of progressively finer stochastic orderings, which they call *inverse stochastic dominance*, and demonstrate that each of these dominance relations implies the same ordering of S-Gini indices (Weymark, MathSocSci 1(1981), 409-430) having a large enough exponent. Inverse stochastic dominance employs the quantile functions just in the same way as usual stochastic dominance employs the distribution functions of two random variables. The present talk develops orderings of multidimensional socio-economic inequality, that are based on generalized Gini social evaluation functions and a multivariate extension of inverse stochastic dominance. As the notions are related to Yaari's dual theory of choice under risk, they are mentioned as *dual stochastic dominance*. Properties of the dominance orderings are investigated and their relation to central (or trimmed) regions of a distribution is pointed out.

10:15-10:45 **Central limit theorem and asymptotics for solutions of kinetic equations**

Eugenio Regazzini (Universita' degli studi di Pavia)

Abstract: On the basis of an analogy pointed out by H. McKean Jr in 1966, a few years ago we started a program which aims at studying the long-time behavior of solutions of some kinetic equations, by means of representations which connect these solutions to probability laws of certain weighted sums of independent and identically distributed random variables. The discovery of the right representation is comparatively simple for the solution of the spatially homogeneous one-dimensional Kac equation. This fact has represented a cornerstone for the first part of the aforesaid program and has produced, concerning the Kac equation, both new results and improvements on existing ones. The second part is concerned with a more difficult problem, concerning the Boltzmann equation for Maxwelliam molecules, i.e., the proof of a McKean's conjecture about the "best" rate of relaxation to equilibrium of its solution. An outline of the main results is given in the present talk.

10:45-11:05 **On the stick-breaking representation for Gibbs-type priors**

Stefano Favaro (Università di Torino e Collegio Carlo Alberto)

Abstract: Random probability measures are the main tool for Bayesian nonparametric inference, given their law acts as a prior distribution. Many well-known priors used in practice admit different, though (indistribution) equivalent, representations. Some of these are convenient if one wishes to thoroughly analyze the theoretical properties of the priors being used, others are more useful in terms of modeling and computation. As for the latter purpose, the so-called stick-breaking constructions certainly stand out. In this talk we focus on the recently introduced class of Gibbs-type priors and provide a stick-breaking representation for it.

11:05-11:30 **Coffee Break**

11:30-12:00 **Chernoff's distribution is log-concave ... and more**

Jon Wellner (University of Washington, Seattle, USA)

Abstract: Chernoff's distribution, which apparently arose first in connection with nonparametric estimation of the mode of a uni-modal density, is the density of the location of the maximum of two-sided Brownian motion minus a parabola. By Groeneboom's switching relation this random variable has the same distribution as the slope at zero of the least concave majorant of two-sided Brownian motion minus a parabola, and it is in this connection that it arises naturally as the limit distribution (up to multiplicative constants) of nonparametric estimators of monotone functions. It was studied further by Daniels and Skryme (1985) and by

Groeneboom (1985), (1989), and computed by Groeneboom and Wellner (2001). It appears ``Gaussian' in shape; and it is natural to conjecture that Chernoff's distribution is log-concave.
In this talk I will give an outline of the proof of log-concavity of Chernoff's distribution, and present graphical evidence in support of the conjecture that Chernoff's distribution is strongly log-concave. This conjecture remains an open problem.

12:00-12:30 **Statistical inference for fractional diffusion processes**

B.L.S. Prakasa Rao (University of Hyderabad, India and Indian Statistical Institute)

Abstract: We introduce some properties of self-similar processes, fractional Brownian motion, fractional diffusion processes and study inference problems for parameters involved in such processes.

12:30-13:00 **A Nonparametric Bayesian Model for Local Clustering***

Peter Mueller (University of Texas, Austin, USA)

Abstract: We propose a nonparametric Bayesian local clustering (NoB-LoC) approach for heterogeneous data. Using genomics data as an example, the NoB-LoC clusters genes into gene sets and simultaneously creates multiple partitions of samples, one for each gene set. In other words, the sample partitions are nested within the gene sets. Inference is guided by a joint probability model on all random elements. Biologically, the model formalizes the notion that biological samples cluster differently with respect to different genetic processes, and that each process is related to only a small subset of genes. These local features are importantly different from global clustering approaches such as hierarchical clustering, which create one partition of samples that applies for all genes in the data set. Furthermore, the NoB-LoC includes a special cluster of genes that do not give rise to any meaningful partition of samples. These genes could be irrelevant to the disease conditions under investigation. Similarly, for a given gene set, the NoB-LoC includes a subset of samples that do not co-cluster with other samples. The samples in this special cluster could, for example, be those whose disease subtype is not characterized by the particular gene set. We provide extensive examples to demonstrate the unique features of the NoB-LoC.

** Joint work with Juhee Lee and Yuan Ji.*

13:00-14:30 Bocconi Buffet

14:30-15:00 **Bayesian Non parametric identication of Pre-versus-Post treatment biomarker effects on Progression free survival***

Michele Guindani (University of Texas, M.D. Anderson Cancer Center, Houston, USA)

Abstract: Recently, much progress has been made in testing the benefit of drugs directed against molecular pathways involved in the development of a disease. However, the identification of biomarkers and compounds capable to accurately explain response to therapy is still a major challenge, especially in oncology. We develop a Bayesian Semiparametric model that enables us to accurately describe the heterogeneity in individual biomarkers' profiles and their association with patients' clinical outcomes. Incidentally, we discuss how changes over time of the biomarkers' profile can be taken into account in the survival regression model by properly redefining a Bayesian Nonparametric estimator of P(X<Y). We discuss the performance of our model in an application to data from patients affected by a form of prostate cancer.

**This is joint work with R. Graziani (Universita' Bocconi), P. Matthew and P. Thall (UT MD Anderson Cancer Center).*

15:00-15:30 **Bayesian nonparametric priors and Markov processes**

Matteo Ruggiero (Università di Torino e Collegio Carlo Alberto)

Abstract: We will review some recent advances on the interplay between certain random probability measures, widely used in Bayesian nonparametric settings for inferential purposes, and specific instances of infinite-dimensional or measured-valued diffusions. Special emphasis will be put on their construction via finite-dimensional Markov chains and diffusions, together with the interpretation of such processes in a biological framework.

15:30-16:00 Coffee Break

16:00-16:20 **More Informative Conditional Density Estimation with Multivariate Covariates**

Sara Wade (Università Bocconi)

Abstract: Flexible conditional density estimation can be achieved by modeling the joint density of the response y and covariate x as a Dirichlet process mixture. Moreover, computations are relatively easy. In this presentation, we examine the predictive performance of these models with an increasing number of covariates and propose a solution based on the Enriched Dirichlet process that overcomes issues encountered by making better use of the information provided by the data. Our proposal maintains a simple allocation rule, so that computations remain relatively simple. Advantages are shown through both predictive equations and examples, including an application to diagnosis Alzheimer's disease.

16:20-16:40 **Bayesian estimation of the discrepancy with misspecified parametric models**

Pierpaolo de Blasi (Università di Torino e Collegio Carlo Alberto)

Abstract: We study a Bayesian semi-parametric model where we have made specific requests about the parameter values to be estimated. The aim is to find the parameter of a parametric family which minimizes a distance to the data generating density and then to find the discrepancy using nonparametric methods. We illustrate how coherent Bayesian updating can proceed given that we are out of the standard Bayes framework. Bayesian updating is performed using MCMC methods and in particular a novel method for dealing with intractable normalizing constants is required. Illustrations using synthetic data are provided.

16:40-17:00 **Free contributions**

17:00-17:30 **Free contributions**

Friday, September 14th

9:45-10:15 **On information distortion in online ratings**

Marco Scarsini (Università LUISS Guido Carli, Roma)

Abstract: The use of reviews has become ubiquitous on the internet, where websites allow users to comment on the products or services they purchased. Typically the review consists of a grade and a written comment. Such grades, or their aggregate statistics, are often the basis for other users' decisions, who in turn may post their reviews. The research question that this paper considers is the following: What does the ultimate distribution of reviews and aggregate statistics represent, given that reviews are posted sequentially rather than simultaneously? We analyze different types of consumer behavior with respect to earlier reports. In particular we focus on two broad classes of behaviors: compensating consumers, who use their rating to attempt to correct an existing evaluation that is perceived as unfair, and conforming consumers, who tend to follow the crowd.

** Joint work with Omar Besbes*

10:15-10:45 **Cross sectional sampling, bias and dependence***

Yosef Rinott (Hebrew University of Jerusalem)

Abstract: We consider a population such as hospital patients, which subjects enter according to certain discrete processes. The population is sampled at a given time (or times), resulting in a cross-sectional sample that includes all those that are present at sampling time. This type of sampling leads to a random sample size, a version of length biased durations of stay (lifetime) in the populations, and dependent observed lifetimes. The goal is to estimate the distribution of lifetime in the population. We discuss various models and methods, parametric and non-parametric, and compare them.

** Joint unfinished work with Micha Mandel.*

10:45-11:05 **Some results on stochastic comparisons of ROC curves***

Chiara Gigliarano (Università Politecnica delle Marche)

Abstract: The main objective of this paper is to propose a novel approach for model comparisons when ROC curves show one or more intersections. We investigate in a theoretical framework the relationship between ROC orderings and stochastic dominance, and we propose new indicators for performance evaluation.

** Joint work with Silvia Figini and Pietro Muliere*

11:05-11:30 Coffee Break

11:30-12:00 **On the estimation of distribution functions in sampling finite populations**

Pierluigi Conti (Universita' La Sapienza, Roma)

Abstract: The estimation of the distribution function of a population is an important problem in sampling finite populations. The existing literature essentially considers the problem of estimating the population distribution function (p.f.d.) at a single point, or at a finite number of points. This talk focuses on the problem of estimating the whole p.d.f.. In many respects, the starting point is close to classical nonparametric statistics, although the approach to inference is based on sampling design. It is shown that the Hajek estimator of the p.d.f., if properly centered and scaled, converges weakly to a Gaussian process with covariance kernel proportional to that of a Brownian bridge. The proportionality factor essentially depends on the sample design. Applications to construction of a confidence band for the p.d.f, and testing for independence of two characters are discussed.

12:00-12:30 **Bayesian Poisson Mixtures for Frequencies of Frequencies***

Luca Tardella (Universita' La Sapienza, Roma)

Abstract: In many scientific contexts observed data can be often expressed as a series of counts. For example in capture-recapture analyses each count reported for all M observed units represent the number of captures occurred during whole experiment. On the other hand in species problems counts represent the number of units observed for each specie while in genetic studies they are associated to the number of active genes, expressed sequence tags or clonetypes. All the statistical information associated to the counts can be summarized by the sufficient statistics of frequencies of distinct counts, the so-called frequencies of frequencies. We show how it is possible to flexibly model the frequencies of frequencies within the framework of nonparametric mixtures of Poisson distribution using a convenient moment-based parameterization of the mixing distribution for which no parametric distribution assumption is made. We compare results from our approach with the most recent crossvalidation-based approach for Poisson compound gamma model proposed in Wang 2010.

** Joint work with Danilo Alunni Fegatelli*

12:30-13:00 **Gene Set Analysis as a Tool for Cross-Platform Integration in Genomics**

Giovanni Parmigiani (Dana Farber Cancer Institute and Harvard School of Public Health, USA)

Abstract: Gene set analysis considers whether genes that form a set from a specific biological standpoint, also behave similarly in a high throughput genomic experiment. This simple cross referencing is very powerful: creative definition of sets has allowed combined analysis and interpretation of very disparate sources of knowledge. In this presentation I will provide a brief review of concepts, our ongoing research on models for gene set analysis, and remaining challenges.

13:00-14:30 Lunch Time

14:30-15:00 **Mixture Importance Splitting. Normalisation and Rare Events**

Nicholas Polson (University of Chicago Booth School of Business)

Abstract: We will discuss Mixture Importance Splitting as a unifying perspective in Monte Carlo estimation of expectations, and as a useful and sometimes necessary variance reduction technique.

15:00-15:30 **Canonical correlations for dependent random measures***

Dario Spanò (University of Warwick, UK)

Abstract: Can Bayes meet Lancaster? In the 1950-60's Sarmanov and Lancaster introduced the method of canonical correlations to build bivariate distributions with fixed marginal laws. Such bivariate distributions are described by transition kernels with fixed orthogonal polynomial eigenfunctions (uniquely determined by the given marginal laws), and are therefore identified by the associated sequence of eigenvalues. In this talk we illustrate how the canonical correlation approach can be employed in an infinite-dimensional context, to construct bivariate random measures when the given marginals are Gamma completely random measures or Dirichlet processes. For such laws, a complete characterization of all possible eigenvalue sequences is possible. The main advantage of this method is all their conditional and joint finite-dimensional distributions admit explicit series representations, and simple algorithms for simulations are available based on dependent Polya urn schemes. For this reason, canonical correlated gamma and Dirichlet processes are appeaing for their potential applications as Bayesian nonparametric priors in a context of partially exchangeable data.

** Based on joint works with Bob Griffiths (Oxford) and Antonio Lijoi (Pavia).*

15:30-16:00 Coffee Break

16:00-16:20 **Bayesian Nonparametric Analysis of reversible Markov Chains***

Lorenzo Trippa (Harvard School of Public Health, USA)

Abstract: We introduce a three-parameter random walk with reinforcement, which generalizes the edge reinforced random walk of Diaconis and Coppersmith, as well as the exchangeable urn scheme that gives rise to the Pitman-Yor process. We use the process to define a prior for reversible Markov chains via de Finetti's theorem for Markov chains. A prior parameter beta smoothly tunes the model between the special cases mentioned and modulates the concentration of the prior, which we demonstrate could be important in applications. We provide a Gibbs sampler for posterior inference. The analysis is applied to the estimation of reversible Markov

models of molecular dynamics simulations. In addition, the extra parameter makes it possible to define the prior on infinite spaces, which allows us to address the problem of species sampling from a reversible Markov chain in a Bayesian nonparametric fashion.

** Joint work with Sergio Bacallado and Stefano Favaro*

16:20-16:40 **Stochastic orderings: the sequel**

Antonietta Mira (Università della Svizzera italiana)

Abstract: Orderings defined on the space of transitions kernels having a specified stationary distribution will be discussed together with the implications that these orderings have when comparing Markov chain Monte Carlo (MCMC) algorithms in terms of their asymptotic efficiency. Insight will be gained to improve the Metropolis-Hastings-Green type of algorithms modifying their corresponding transition kernels by delaying the rejection step. The resulting Delayed Rejection MCMC kernels will be shown to dominate the corresponding original transition kernels in terms of the efficiency ordering introduced.

16:40-17:00 **Simpson's Paradox in Survival Models**

Clelia di Serio (Università Vita-Salute San Raffaele, Milano)

Abstract: Accounting for unobservable heterogeneity in the population strata is a very hard concept in biostatistics. In the context of survival analysis it is possible that increasing the value of a covariate X has a beneficial effect on a failure time, but this effect is reversed when conditioning on any possible value of another covariate Y. When studying causal effects and influence of covariates on a failure time, this state of affairs appears paradoxical and raises questions about the real effect of X. Situations of this kind may be seen as a version of Simpson's paradox. The introduction of a time variable makes the paradox more interesting and intricate: it may hold conditionally on a certain survival time, i.e. on an event of the type {T>t} for some but not all t, and it may hold only for some range of survival times. An application of Simpson paradox to prostate cancer data is illustrated.

Evening **Social Dinner**

Saturday, September 15th

10:10-10:30 **Proper and non-informative conjugate priors for exponential family models***

Eduardo Gutiérrez Peña (IIMAS, UNAM, Mexico D.F.)

Abstract: Exponential families constitute an important class of probability models that occur, in one form or another, as part of more complex models widely used in applied statistics such as generalized linear models. These families are related to the notion of sufficiency and can also be motivated as a set of solutions to certain maximum entropy problems. On the other hand, conjugate distributions play a relevant role in the Bayesian approach to parametric inference. One of the main features of such families is that they are closed under sampling, but a conjugate family often provides prior distributions which are tractable in various other respects. In particular, conjugate families for exponential family models are themselves exponential families, and so can also be regarded as solutions to maximum entropy problems.
Maximum entropy (or, more generally, minimum relative entropy) is known to be closely related to the decision theoretical problem of minimizing a worst-case expected loss. In the context of exponential families, it is also noteworthy that, under certain conditions, Jeffreys' and other non-informative priors –including some forms of 'unbiased' priors– can be obtained as suitable limits of conjugate distributions. Moreover, there exists an interesting duality between unbiased estimators and optimal Bayes estimators that minimize expected risk. In this talk we discuss these various concepts and highlight the relationship between them.

** Joint work with Manuel Mendoza*

10:30-11:00 **Uniformly Most Powerful Bayesian Tests**

Valen Johnson (University of Texas, Anderson Cancer Center, Houston, USA)

Abstract: Uniformly most powerful tests are statistical hypothesis tests that provide the greatest power against a fixed null hypothesis among all possible tests of a given size. In this talk, I extend the notion of uniformly most powerful tests to the Bayesian setting by defining a uniformly most powerful Bayesian test to be a test which maximizes the probability that the Bayes factor in favor of the alternative hypothesis exceeds a given threshold. Like their classical counterpart, uniformly most powerful Bayesian tests are most easily defined in one-parameter exponential family models, although I demonstrate that extensions outside of this class are possible. I also show how the connection between uniformly most powerful tests and uniformly most powerful Bayesian tests can be used to provide an approximate calibration between p-values and Bayes factors. Finally, I discuss issues of regarding the strong dependence of both the resulting Bayes factors and p-values on sample size.

11:00-11:30 Coffee Break

11:30-12:00 **The Measurement of Multi-group Dissimilarity and Related Orders***

Claudio Zoli (Università di Verona)

Abstract: The analysis of multi-group segregation, socioeconomic mobility evaluation, equalization of opportunity and discrimination involve the measurement of the dissimilarity between sets of at least two probability distributions, made conditional on a partition of the population in groups, and defined over a finite number of classes. We analyze the dissimilarity partial order, a general criterion for ranking sets of distributions according to the degree of dissimilarity between the elements of these sets. Firstly, we axiomatically characterize the dissimilarity order, using the dissimilarity reducing transfers/exchanges of population masses both across classes and/or groups. Then we show that these operations identify an equivalent representation for the dissimilarity order based on Dahl's matrix majorization criterion [1999, Matrix Majorization, Lin. Alg. App. 288:53-73]. Secondly, we show that the dissimilarity order is empirically testable. If classes are permutable, we prove that the dissimilarity order is equivalent to the ranking of sets of distributions produced by the inclusion of the Zonotopes associate to the sets. Conversely, if classes are ordered, the test resorts on a finite number of Lorenz majorization comparisons among groups conditional proportions, performed at different cumulation stages of the overall population. The relation with the related concepts of inequality, segregation and discrimination is also discussed.

** Joint work of Francesco Andreoli and Claudio Zoli*

12:00-12:30 **Reinforced urns: a biased overview**

Piercesare Secchi (Politecnico di Milano)

Abstract: A personal tribute to Pietro Muliere's legacy on reinforced urns and their impact on statistical modeling.

12:30-13:00 **Closing**