Draft 11th of July 2013 Fabrice Gamboa
fabrice.gamboa@math.univ-toulouse.fr

Summer school in Mathematical Statistics in Vietnam July 2014

Aim and scope

The aim of the school is to give an overview on mathematical statistics and its applications. This set of postgraduate courses will provide a sharp introduction to many hot topics actually investigated in the statistical paradigm. The aim is mainly twofold. First, for Mathematical students it will give the flavor of what are modern mathematical statistics. Secondly, it will be the springboard for a huge scientific cooperation in Statistics between *Vietnam's Universities* and *l'Institut de Math´ematiques de Toulouse* and more generally with other statistical French labs. We expect around 50 Vietnamese postgraduate students and researchers. All the courses are 40 hours including practical tutorials on computer. The school may held in Ho Chi Minh City during the whole month of August 2014. There will be 5 courses at the same time during the whole month of August. Two lecturers will be associated to each courses. The prerequisite to attend these courses will be the good knowledge of the following items :

Probability

- Probability space, Borel Cantelli Lemma.
- Law, density, repartition function, moments.
- Convergences : in probability, almost surely, weakly.
- *L*2 space. Gaussian calculus.
- Law of large numbers, central limit Theorem.

Statistics

- Empirical estimation, basic properties of estimator.
- Likelihood, Fisher information, Cram´er Rao bound.
- Gaussian linear model.
- Exponential model.
- Introduction to test theory.

In order to prepare the students that does not handle all the prerequisite items, a two weeks intensive pregraduate course on these topic may be programmed in April 2014. This preparing course will be of 50 hours, including practical tutorials on computer. Two Professors of the french team will come two weeks in Vietnam for this spring training courses.

Possible courses for the summer school

This is a first list of possible courses and possible professors.

1. Large deviations, concentration inequalities and statistical applications
   Bernard Bercu (Bordeaux) Bernard.Bercu@math.u-bordeaux1.fr Fabrice Gamboa (Toulouse) fabrice.gamboa@math.univ-toulouse.fr Thierry Klein (Toulouse) thierry.klein@math.univ-toulouse.fr

2. Sensitivity analysis and model reduction for computer experiments
   Fabrice Gamboa (Toulouse) fabrice.gamboa@math.univ-toulouse.fr
   Thierry Klein (Toulouse) thierry.klein@math.univ-toulouse.fr
   Agn`es Lagnoux-Renaudie (Toulouse)lagnoux@math.univ-toulouse.fr

3. Introduction to model selection and applications.
   B´eatrice Laurent-Bonneau (Toulouse) Beatrice.Laurent@insa-toulouse.fr
   Cl´ement Marteau (Toulouse) clement.marteau@math.univ-toulouse.fr

4. Introduction to non parametric minimax theory
   B´eatrice Laurent-Bonneau (Toulouse) Beatrice.Laurent@insa-toulouse.fr
   Cl´ement Marteau (Toulouse) clement.marteau@math.univ-toulouse.fr

5. Introduction to particle methods and importance sampling/splitting methods
   Pierre del Moral (Bordeaux) Pierre.Del Moral@inria.fr
   Agn`es Lagnoux-Renaudie (Toulouse)lagnoux@math.univ-toulouse.fr

6. Continuous and discrete martingales. Statistical applications. Bernard Bercu (Bordeaux) Bernard.Bercu@math.u-bordeaux1.fr Jean-Yves Dauxois (Toulouse)Jean-Yves.Dauxois@insa-toulouse.fr

7. Unidimensional and multivariate time series analysis with applications
   Xavier Bressaud (Toulouse) bressaud@math.univ-toulouse.fr
   Beno^ıt Truong Van (Toulouse)Benoit.truong-van@insa-toulouse.fr

Lecturers

Bernard Bercu (Bordeaux) Bernard.Bercu@math.u-bordeaux1.fr
Xavier Bressaud (Toulouse) bressaud@math.univ-toulouse.fr
Pierre del Moral (Bordeaux) Pierre.Del Moral@inria.fr
Jean-Yves Dauxois (Toulouse)Jean-Yves.Dauxois@insa-toulouse.fr
Fabrice Gamboa (Toulouse) fabrice.gamboa@math.univ-toulouse.fr
Thierry Klein (Toulouse) thierry.klein@math.univ-toulouse.fr
Agn`es Lagnoux-Renaudie (Toulouse)lagnoux@math.univ-toulouse.fr
B´eatrice Laurent-Bonneau (Toulouse) Beatrice.Laurent@insa-toulouse.fr
Cl´ement Marteau (Toulouse) clement.marteau@math.univ-toulouse.fr
Beno^ıt Truong Van (Toulouse)Benoit.truong-van@insa-toulouse.fr

Summary of the different courses

1. **Large deviations, concentration inequalities and statistical applications.**
   The aim of this course is to give a first and deep insight into the theory of large deviations and concentration of measures. A first part of the course will be devoted to the develop- ment of the general concepts and theorems. In a second part, we will provided some applications in statistics. Some practical works on computer will illustrate the theory.

- Cramer theorem on R. Sharp large deviations.
- Bennet and Hoeffding inequalities.
- Abstract large deviations.
- Concentration of Gaussian and bounded supported measures.
- Statistical applications : confidence sets and asymptotic properties of test.

Short bibliography

[1] *JA Bucklew*. Large deviation techniques in decision, simulation, and estimation. Wiley Series in Pro- bability and Mathematical, 1990.

[2] *A Dembo, O Zeitouni* . Large deviations techniques and applications. Volume 38 of Stochastic Mo- delling and Applied Probability Springer-Verlag, Berlin, 2010.

[3] *M Ledoux* . The concentration of measure phenomenon. American Mathematical Society, Providence, RI, 2001.

2. **Sensitivity analysis and model reduction for computer experiments.**
   Many mathematical models encountered in applied sciences involve a large number of poorly-known parameters as inputs. It is important for the practitioner to assess the impact of this uncertainty on the model output. An aspect of this assessment is sensiti- vity analysis, which aims to identify the most sensitive parameters, that is, parameters having the largest influence on the output. In global stochastic sensitivity analysis (see for example [2] and [3] and references therein) the input variables are assumed to be inde- pendent random variables. Their probability distributions account for the practitioner's belief about the input uncertainty. This turns the model output into a random variable, whose total variance can be split down into different partial variances (this is the so-called Hoeffding decomposition see [1]). Each of these partial variances measures the uncertainty on the output induced by each input variable uncertainty. By considering the ratio of each partial variance to the total variance, we obtain a measure of importance for each input variable that is called the Sobol index or sensitivity index of the variable ([4]) ; the most sensitive parameters can then be identified and ranked as the parameters with the largest Sobol indices. In this course, we will present some statistical properties of the estimators of the Sobol indices.

Short bibliography

[1] *W. Hoeffding*. A class of statistics with asymptotically normal distribution. Ann. Math. Statistics, 19 : 293–325, 1948.

[2] *Jeremy E. Oakley and Anthony O'Hagan*. Probabilistic sensitivity analysis of complex mo- dels : a Bayesian approach. J. R. Stat. Soc. Ser. B Stat. Methodol., 66(3) :751–769, 2004.

[3] *A. Saltelli, K. Chan, and E.M. Scott*. Sensitivity analysis. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 2000.

[4] *I. M. Sobol*. Sensitivity estimates for nonlinear mathematical models. Math. Modeling Comput. Ex- periment, 1(4) :407–414 (1995), 1993.

3. **Introduction to model selection and applications.**

The goal of this course is to provide an overview of the model selection topic. Given data, model selection aims at finding the most appropriate model that could explain the data, according to the behavior of additional explanatory variables.

In a first time, we will deal with the General Linear Model and in particular in a multivariate regression framework. In this context, we will recall the main statistical tool that allow to provide some inference on the data : Fisher's test, contruction of confidence bounds,.... Then we will introduce the backward and forward regression algorithms and the well-known Cp, AIC and BIC criteriums. Practical experiments on computer (with the statistical software R) will be provided.

In a second time, we will investigate the theoretical properties of recent model selection algorithms in a general context. In particular, we will prove that the behaviour of theses methods can be controlled in a satisfying way.

Short bibliography

[1] *A. Barron, L. Birg´e and P. Massart*. Risk bounds for model selection via penalization. Probability Theory and Related Fields, 113 :301-413 (1999).

4. **Introduction to non parametric minimax theory.** In this course our aim is to present well-known non-parametric models, to construct in each case appropriate estimators and to investigate related theoretical properties. To this end, the course will decomposed in three different parts :

- The gaussian white noise model : a short historical background will be provided with re- lated applications. Some attention will be payed to the construction of linear estimators (tests) and to the study of related minimax properties.
- The regression model : several kind of estimators will be presented in this context (kernel based, splines, estimators based on histograms...).
- the density model : kernel methods are at the earth of non-parametric inference in such a model. We will investigate both theoretical and numerical properties of these methods. We will also discuss the error-in-variable that allow to take into account measurements error in our models. The underlying problem appears to be an inverse (convolution) problem and require appropriates methods.

Short bibliography

[1] *A.B. Tsybakov*. Introduction to nonparametric estimation. Springer series in statistics (2008).

5. **Introduction to particle methods and importance sampling/splitting methods**. The analysis of rare events is of great importance in many fields because of the risk associated to the event. Their probabilities are usually less than 10^{-9}. One can use many ways to study them : the first one is statistical analysis, based on the standard extreme value distributions but needs a long observation period (see Aldous [1]), the second one is modelisation which leads to estimate the rare event probability either by analytical approach (see Sadowsky [5]), or by simulation. In this course, we focus on the simulation approach based on Monte-Carlo method. Nevertheless crude simulation is impracticable for estimating such small probabilities : to estimate proba- bilities of order 10^{-10} with acceptable confidence would require the simulation of at least one thousand billion events

(which corresponds to the occurrence of only one hundred rare events). To overcome these limits, fast simulation techniques are applied. In particular, importance sam- pling (IS) is a refinement of Monte-Carlo methods. The main idea of IS ([4]) is to make the occurrence of the rare event more frequent. More precisely IS consists in selecting a change of measure that minimizes the variance of the estimator. Another method is called splitting. The basic idea of splitting ([3]) is to partition the space-state of the system into a series of nested subsets and to consider the rare event as the intersection of a nested sequence of events. When a given subset is entered by a sample trajectory, random retrials are generated from the ini- tial state corresponding to the state of the system at the entry point. More refined versions of splitting as particles systems ([2]) or RESTART ([6]) have been introduced in the last decades.

Short bibliography

[1] *David Aldous*. Probability approximations via the Poisson clumping heuristic, volume 77 of Applied Mathematical Sciences. Springer-Verlag, New York, 1989.

[2] *Pierre Del Moral* Feynman-Kac formulae. Probability and its Applications (New York). Springer- Verlag, New York, 2004. Genealogical and interacting particle systems with applications.

[3] *Agn`es Lagnoux*. Rare event simulation. Probability in the Engineering and Informational Sciences, 20(1) : 45–66, 2006.

[4] *Bernard Lapeyre, Etienne Pardoux, and R´emi Sentis*. M´ethodes de Monte-Carlo pour les ´equations de transport et de diffusion. Math´ematiques & Applications [Mathematics & Ap- plications], 29. Springer-Verlag, Berlin, 1998.

[5] *John S. Sadowsky*. On Monte Carlo estimation of large deviations probabilities. Ann. Appl. Probab., 6(2) : 399–422, 1996.

[6] *Manuel Vill´en-Altamirano and Jos´e Vill´en-Altamirano*. Analysis of restart simulation : Theoretical basis and sensitivity study. European Transactions on Telecommunications, 13 n4 :373–385, 2002.

6. Continuous and discrete martingales. Statistical applications.

The goal of the first part of this course is to provide an overview of historical and recents results on concentration inequalities for sums of independent random variables and for martingales. First of all, we shall deal with classical concentration inequalities for sums of independent random variables such as the famous Hoeffding, Bennett, Bernstein and Talagrand inequalities. Further results and improvements will also be provided such as the missing factors in thoses inequalities. The second part of the course concerns concen- tration inequalities for martingales such as Azuma-Hoeffding, Freedman and De la Pena inequalities. Several extensions will also be provided. Finally, we shall deal with applica- tions of concentration inequalities in probability and statistics.

The second part of this lecture will introduce part of the theory of continuous time martingales with aim to present their use in Statistic, particularly in Survival Analysis. The content should not be far from the following.

- Stochastic process, Filtration, Stopping Time, Localization, Stochastic Stieltjes Inte- gral.
- (sub, super) Martingale, Squarre Integrable Martingale, Stopping Theorems, Predic- table Process, Doob-Meyer Decomposition, Predictable Variation Process, Local Mar- tingale.
- Martingale Transform, Predictable Covariation Process.
- Lenglart Inequality, Rebolledo Central Limit Theorem for Martingales.
- Counting Process, Compensator, Predictable Variation Process, Intensity, Innovation Theorem, Stopped and Filtered Counting Process, Censoring.
- Nonparametric Statistic with Counting Processes and Applications In Survival Analy-

sis, Nelson-Aalen and Kaplan-Meier Estimators, Large Sample Behavior.

7. Unidimensional and multivariate time series analysis with applications

To be announced.