

# Journées Statistiques du Sud 2009

Site de Porquerolles, 17-19 juin 2009

*L'accueil des participants aura lieu le mardi 16 juin 2009 à partir de 18h00 sur le site IGESA de Porquerolles.*

## MERCREDI 17 JUIN 2009 :

- 8h30-9h00 : Accueil des participants et présentation des journées.
- 9h00-10h30 : **Oleg Lepski** (Aix-Marseille I)  
*Part I : Universal selection rule in non-parametric estimation.*
- 10h30-11h00 : *Break*
- 11h00-12h30 : **Gilles Blanchard** (Fraunhofer First)  
*Part I : Non-asymptotic adaptive control of the family-wise error rate in multiple testing.*

## *LUNCH*

- 14h00-15h00 : **Olivier Lopez** (Paris VI)  
*Non-parametric model check for parametric mean-regression under random censoring.*
- 15h00-16h00 : **Jean-Michel Zakoian** (CREST-Paris)  
*Recent results on the estimation and prediction of conditionally heteroskedastic models.*
- 16h00-16h30 : *Break*
- 16h30-17h30 : **Patricia Reynaud-Bouret** (CNRS-Nice)  
*Adaptive estimation for Hawkes processes; application to genome analysis.*

**JEUDI 18 JUIN 2009 :**

- 9h00-10h30 : **Ingrid Van Keilegom** (Louvain)  
*Part I : The Empirical Likelihood Method in Regression.*
- 10h30-11h00 : *Break*
- 11h00-12h30 : **Gilles Blanchard** (Fraunhofer First)  
*Part II : False Discovery Rate control in multiple testing: sufficient conditions and adaptivity.*

*LUNCH*

- 14h00-15h00 : **Elodie Brunel** (Montpellier II)  
*Penalized contrast estimation of cumulative distribution functions in survival models.*
- 15h00-16h00 : **Erwan Le Pennec** (Paris VII)  
*Maxisets for model selection.*
- 16h00-16h30 : *Break*
- 16h30-17h30 : **Christophe Pouet** (Aix-Marseille I)  
*Test on components of densities mixture.*

**VENDREDI 19 JUIN 2009 :**

- 8h30-10h00 : **Ingrid Van Keilegom** (Louvain)  
*Part I : The Empirical Likelihood Method in Regression.*
- 10h00-10h30 : *Break*
- 10h30-12h00 : **Oleg Lepski** (Aix-Marseille I)  
*Part II : Uniform bounds for norms of sums of independent random functions.*

*LUNCH*

**Oleg Lepski**  
Aix-Marseille I University

*joint work with A. Goldenshluger, Haifa University.*

**PART I : Universal selection rule in non-parametric estimation.**

This part is devoted to the discussion of a new approach to nonparametric estimation which is based on the selection from a given family of *linear estimators*. The important issue of our methodology is its application in various statistical settings since there is no restrictions related to the statistical model.

Let  $(\mathcal{X}^{(n)}, \mathfrak{B}^{(n)}, \mathbb{P}_f^{(n)}, F \in \mathbb{F})$  be a family of statistical experiments generated by an observation  $X^{(n)}$ . It means that  $\mathfrak{B}^{(n)}$  is  $\sigma$ -algebra generated by random element  $X^{(n)}$  and, necessarily, the probability law of  $X^{(n)}$  belongs to the family  $(\mathbb{P}_f^{(n)}, f \in \mathbb{F})$ .

Let  $\mathcal{D}$  be the open interval in  $\mathbb{R}^d$ ,  $d \geq 1$ , let  $\mathbb{F}$  be the set of borel functions  $f : \mathcal{D} \rightarrow \mathbb{R}$  and let  $\mathbf{m}$  be the  $\sigma$ -finite measure on  $\mathcal{D}$ . Our goal is to estimate the function  $f$ . To avoid the discussion of boundary effects we are interesting in estimating  $f$  on  $\mathcal{D}_0$ , where  $\mathcal{D}_0$  is an open subinterval of  $\mathcal{D}$ . As an estimator of  $f$  we understand any  $\mathfrak{B}^{(n)}$ -measurable mapping,  $\hat{f} : \mathcal{X}^{(n)} \times \mathcal{D} \rightarrow \mathbb{F}_0$ , where  $\mathbb{F}_0 \supseteq \mathbb{F}$  is a separable linear metric space of functions defined on  $\mathcal{D}$  and acting to  $\mathbb{R}$ . With any estimator we associate the risk

$$\mathcal{R}_n^{(\ell)}[\hat{f}; f] = \left( \mathbb{E}_f^{(n)} \left[ \ell(\hat{f} - f) \right]^q \right)^{\frac{1}{q}},$$

where  $\ell : \mathbb{F}_0 \rightarrow \mathbb{R}_+$  is a **semi-norm** and  $q > 0$  is a given number. The problem is to estimate  $f$  from observation  $X^{(n)}$  with small risk  $\mathcal{R}_n^{(\ell)}[\hat{f}; f]$  at least for large  $n$ .

*We say that the estimator  $\hat{f}(x)$  is linear if  $\exists K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  such that*

$$\mathbb{E}_f[\hat{f}(x)] = \int_{\mathcal{D}} K(t, x) f(t) \mathbf{m}(dt), \quad \forall f \in \mathbb{F}, \quad \forall x \in \mathcal{D}.$$

Thus, the linear estimator is the estimator whose mathematical expectation is the linear functional of underlying function  $f$ . Let  $\mathcal{D}_1$  be an open interval such that  $\mathcal{D}_0 \subseteq \mathcal{D}_1 \subseteq \mathcal{D}$ .

*Any function  $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$  such that*

$$\begin{aligned} \int_{\mathcal{D}} K(t, x) \mathbf{m}(dt) &= 1, \quad \forall x \in \mathcal{D}_1; \\ \text{supp}(K(\cdot, x)) &\subseteq \mathcal{D}_1, \quad \forall x \in \mathcal{D}_0, \end{aligned}$$

*will be called  $\mathcal{D}_1$ -weight and let  $\mathfrak{K}(\mathcal{D}_1)$  be the set of all  $\mathcal{D}_1$ -weights.*

We endow  $\mathfrak{K}(\mathcal{D}_1)$  with the operation " $\otimes$ ":  $\forall K_1, K_2 \in \mathfrak{K}(\mathcal{D}_1)$

$$[K_1 \otimes K_2](\cdot, \cdot) = \int_{\mathcal{D}_1} K_1(\cdot, y) K_2(y, \cdot) \mathbf{m}(dy).$$

and we say that  $K_1 \in \mathcal{K}$  and  $K_2 \in \mathcal{K}$  commute if

$$[K_1 \otimes K_2] \equiv [K_2 \otimes K_1].$$

We say that  $\mathcal{K} \in \mathfrak{K}(\mathcal{D}_1)$  is commutative weight system if any pair of its elements commute.

Let  $\mathcal{K}$  be a commutative weight system and let  $\mathcal{L}_{\mathcal{K}} = \mathcal{K}^{\otimes} \cup \mathcal{K}$ , where

$$\mathcal{K}^{\otimes} = \{L : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} : L = K \otimes K', K, K' \in \mathcal{K}\}.$$

Suppose that  $(\mathcal{X}^{(n)}, \mathfrak{B}^{(n)}, \mathbb{P}_f^{(n)}, f \in \mathbb{F})$  is  $\mathcal{L}_{\mathcal{K}}$ -experiment, i.e a linear estimator is defined for any function belonging to  $\mathcal{L}_{\mathcal{K}}$ .

Let  $\mathcal{F}_{\mathcal{K}} = \{\hat{f}_K, K \in \mathcal{K}\}$  be the collection of linear estimators generated by  $\mathcal{K}$ . We propose the estimator, say  $f^*$ , for the underlying function  $f$  whose construction is based on data-driven selection from the family  $\mathcal{F}_{\mathcal{K}}$ , namely

$$f^* = \hat{f}_{\hat{K}}, \quad \hat{K} = \hat{K}_{X^{(n)}} \in \mathcal{K}.$$

We also establish the explicit upper bound for  $\mathcal{R}_n^{(\ell)}[f^*; f]$  for given  $f \in \mathbb{F}$  and  $n \in \mathbb{N}^*$  in the case of an arbitrary semi-norm  $\ell$ .

The remarkable property of our selection rule is that under rather mild technical assumptions it can be applied to any commutative weight system.

The construction of our selection rule involves the finding of the uniform bounds for rather general stochastic objects. Their descriptions and corresponding results are discussed in the second part of the talk.

## **PART II : Uniform bounds for norms of sums of independent random functions.**

In this part of the talk we present several results related to the upper majorants for the norm of random functions of special type.

Let  $(\mathcal{T}, \mathfrak{T}, \tau)$  and  $(\mathcal{X}, \mathfrak{X}, \varkappa)$  be  $\sigma$ -finite spaces,  $\mathcal{X}$  be a Banach space and let  $(\Omega, \mathfrak{A}, \mathbb{P})$  be a complete probability space.

Let  $X$  be a  $\mathcal{X}$ -valued random element defined on  $(\Omega, \mathfrak{A}, \mathbb{P})$  and having the density  $f$  with respect to measure  $\varkappa$ . Let also  $\varepsilon$  be real random variable defined on the same probability space which is independent of  $X$  and has symmetric distribution.

For any  $(\mathfrak{T} \times \mathfrak{X})$ -measurable function  $w$  on  $\mathcal{T} \times \mathcal{X}$  and for any  $t \in \mathcal{T}$ ,  $n \in \mathbb{N}^*$  we define

$$\xi_w(t) = \sum_{i=1}^n [w(t, X_i) - \mathbb{E}w(t, X)], \quad \eta_w(t) = \sum_{i=1}^n w(t, X_i)\varepsilon_i,$$

where  $(X_i, \varepsilon_i), i = \overline{1, n}$ , are independent copies of  $(X, \varepsilon)$ .

Put, for  $1 \leq s < \infty$ ,

$$\|\xi_w\|_{s,\tau} = \left[ \int |\xi_w(t)|^s \tau(dt) \right]^{\frac{1}{s}}, \quad \|\eta_w\|_{s,\tau} = \left[ \int |\eta_w(t)|^s \tau(dt) \right]^{\frac{1}{s}}$$

and let  $W$  be a given set of  $(\mathfrak{Y} \times \mathfrak{X})$ -measurable functions.

Let  $\Psi_w$  be either  $\xi_w$  or  $\eta_w$ . In this paper we will be interested in finding of *non-random* function on  $W$ , say  $U_\Psi(w)$ , which would be *uniform* upper bound for  $\|\Psi_w\|_{s,\tau}$  in the sense that

$$\mathbb{P} \left\{ \sup_{w \in W} \left[ \|\Psi_w\|_{s,\tau} - u C^*(y) U_\Psi(w) \right] \geq 0 \right\}$$

is small and tends to zero as  $n \rightarrow \infty$  for any fixed  $y > 0$ . Here  $C^*(\cdot)$  is the *given* linear function, and the constant  $u \geq 1$  typically completely determined by  $W$  and often  $u = 1$ .

In fact we want to bound from above the latter probability as well as

$$\mathbb{E} \left( \sup_{w \in W} \left[ \|\Psi_w\|_{s,\tau} - u C^*(y) U_\Psi(w) \right] \right)_+^q, \quad q \geq 1.$$

We provide with explicit expression for  $U_\Psi$  which is completely determined by  $w$ ,  $f$  and  $s$ . We will also show that in the case  $\Psi_w = \eta_w$  the corresponding uniform bound heavily depends on the moment's conditions imposed on the distribution of  $\varepsilon$ .

Another problem arising in the context of the applications of the obtained results in mathematical statistics consists in finding of a bound independent of the density  $f$ . Sometimes this dependence on  $f$  is not crucial because the function  $f$  is supposed to be known. The typical example is the regression model where the random function  $\eta_w$  appears.

There exists, however, the problems where the situation is completely different. One of them is the estimation of unknown multivariate density from *i.i.d.* observations, where  $\xi_w$  appears. In this case  $\xi_w$  can be treated as stochastic part of a linear estimator corresponding to the weight function  $w$ . A uniform non-random bound is used in the selection rule, discussed in Part I, allowing to derive the estimator from a given family of linear estimators. It is clear that the use of a bound dependent on the unknown parameter is impossible for this purpose. To overcome this difficulty we propose a **random** uniform bound, say  $\hat{U}_s(w)$ , whose construction is based only on the sequence  $X_1, \dots, X_n$ , and establish corresponding inequality for

$$\mathbb{E} \left( \sup_{w \in W} \left[ \Psi_w - 2u C^*(y) \hat{U}_s(w) \right] \right)_+^q, \quad q > 0.$$

The constant 2 here can be replaced by any other strictly larger than 1.

The obtained result, together with approach developed in Part I is sufficient to establish very general *oracle inequality* in the context of multivariate density estimation. In particular, it allows to solve completely the problem of *bandwidth selection* for the risks described by  $\mathbb{L}_s$ -norm. The solution was known only for  $s = 1$  and it was obtained by use of absolutely different technique. It allows also to construct the *adaptive* with respect to anisotropic Sobolev classes estimator. This problem was solved only for  $s = 2$  and  $s = \infty$ .

**Gilles Blanchard**  
Fraunhofer First

**PART I : Non-asymptotic adaptive control of the family-wise error rate in multiple testing.**

I will review some (classical and recent) results on non-asymptotical control of the family-wise error rate in multiple testing and adaptivity to (1) the unknown proportion  $\pi$  of true null hypotheses and (2) the unknown correlation structure of the test statistics. These results revolve around the idea of exact tests and step-down methods.

**References:**

Romano, J.P., Wolf, M.: Exact and approximate stepdown methods for multiple hypothesis testing. *JASA* 100(469) (2005) 94- 108.  
L. Wasserman and K. Roeder. Weighted hypothesis testing. Technical report, Dept. of statistics, Carnegie Mellon University, 2006. ArXiv preprint math/0604172v1.  
S. Arlot, G. Blanchard, and E. Roquain. Some non-asymptotic results on resampling in high dimension, I: confidence regions and II : multiple tests, 2009. To appear in the *Annals of Statistics*.

**PART II : False Discovery Rate control in multiple testing: sufficient conditions and adaptivity.**

I will present sufficient conditions ensuring False Discovery Rate control in multiple testing, whose scope allows to recover in an unifying way a number classical methods (such as the step-up, step-down, step-up-down) with extensions. Based on this approach I will discuss the question of adaptivity to the proportion  $\pi$  of true null hypotheses, and if the time permits adaptivity to the alternate distribution, drawing some links with the theory of classification.

**References:**

G. Blanchard and E. Roquain. Two simple sufficient conditions for fdr control. *Electron. J. Stat.*, 2 : 963-992, 2008.  
G. Blanchard and E. Roquain. Adaptive FDR control under independence and dependence, 2009. arXiv :math.ST/07070536v3.  
H. Finner, R. Dickhaus, and M. Roters. On the false discovery rate and an asymptotically optimal rejection curve. *Ann. Statist.*, 37(2) : 596-618, 2009.  
C. Scott and G. Blanchard, “Novelty detection: Unlabeled data definitely help,” D. van Dyk and M. Welling, Eds., *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS) 2009*, JMLR: W&CP 5, 464-471.  
Operating Characteristics and Extensions of the False Discovery Rate Procedure  
C. Genovese and L. Wasserman *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, Vol. 64, No. 3 (2002), 499-517.

Ingrid Van Keilegom  
Louvain University

## The Empirical Likelihood Method in Regression.

In this course, the following topics will be discussed :

1. Basic principles of empirical likelihood (E.L.);
2. E.L. in parametric regression;
3. E.L. in nonparametric regression;
4. E.L. in semiparametric regression;
5. E.L. in regression with missing data;
6. E.L. in regression with censored data;
7. E.L. based goodness-of-fit tests.

For each of the above topics, a summary of the literature will be given, and the most important contributions will be explained in detail. A comparison with more classical inference methods (like those based on asymptotic normality or bootstrap) will also be given, in order to understand better the pros and contras of the empirical likelihood method.

### References:

- Chen, S.X. and Van Keilegom, I. (2009). A goodness-of-fit test for parametric and semiparametric models in multiresponse regression. *Bernoulli* (to appear).
- Chen, S.X. and Van Keilegom, I. (2009). A review on empirical likelihood methods for regression (submitted).
- DiCiccio, T., Hall, P. and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist.*, **19**, 1053-1061.
- Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *Internat. Statist. Review*, **58**, 109-127.
- Hjort, N.L., McKeague, I.W. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.*, **37**, 1079-1115.
- Owen, A.B. (1990). Empirical likelihood confidence regions. *Ann. Statist.*, **18**, 90-120.
- Owen, A. (2001). *Empirical Likelihood*. Chapman and Hall, New York.
- Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300-325.

Olivier Lopez (Paris VI)

**Nonparametric model check for parametric mean-regression under random censoring.**

We consider the problem of nonparametric testing of the null hypothesis,

$$H_0 : \exists \theta_0 \in \Theta \subset \mathbb{R}^k \quad \text{such that} \quad E[Y|X] = f(\theta_0, X),$$

where  $f$  is a known function,  $X \in \mathbb{R}^d$ , and  $Y \in \mathbb{R}$  a randomly right-censored variable. This means, introducing a censoring random variable  $C$ , that we observe an i.i.d. sample  $(T_i, \delta_i, X_i)_{1 \leq i \leq n}$  where

$$\begin{aligned} T_i &= \inf(Y_i, C_i), \\ \delta_i &= \mathbf{1}_{Y_i \leq C_i}. \end{aligned}$$

We first consider the case where the censoring variable is independent from  $(Y, X)$ . In this framework, we obtain consistency of our test procedures against local alternatives. We generalize our procedure to the case where  $C$  is independent from  $Y$  conditionally to  $X$ .

**References:**

- Horowitz, J.L. & Spokoiny, V.G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* 69, 599-631.
- Lopez, O. & Patilea, V. (2009) Nonparametric lack-of-fit tests for parametric mean-regression models with censored data. *Journal of Multivariate Analysis* 100, 210-230.
- Stute, W., González-Manteiga, W. & Sánchez-Sellero, C. (2000). Nonparametric model checks in censored regression. *Comm. Statist. Theory Methods* 29, 1611-1629.
- Zheng, J.X. (1996). A consistent test of functional form via nonparametric estimation techniques. *J. Econometrics* 75, 263-289.

**Jean-Michel Zakoian** (CREST and Lille III)

**Recent results on the estimation and prediction of conditionally  
heteroskedastic models.**

We start by recalling the asymptotic properties of the Gaussian quasi-maximum likelihood estimator (Q.M.L.E.) in GARCH models. The consistency and asymptotic normality hold under mild conditions, including the strict stationarity of the observed process, the existence of fourth-order moments for the strong white noise driving the dynamics and the non nullity of the volatility coefficients. When the moment condition on the noise is not satisfied, the Q.M.L.E. remains consistent but may have a non standard asymptotic distribution. When the parameter stands at the boundary of the parameter space, the asymptotic normality is also in failure: the asymptotic distribution is obtained as the projection of a Gaussian vector on the local parameter space. Based on these results, we consider the problem of optimal prediction of powers, or logarithms, of the absolute process. A standard procedure for estimating this prediction is to estimate the volatility by Gaussian Q.M.L. in a first step, and then use empirical means based on rescaled innovations in a second step. We suggest an alternative one-step procedure, based on an appropriate non-Gaussian Q.M.L. estimation of the model. The performances of the two approaches are compared.

**References:**

- Berkes, I., Horváth, L. and P. Kokoszka (2003) GARCH processes: structure and estimation. *Bernoulli* 9, 201-227.
- Berkes, I. and L. Horváth (2004) The efficiency of the estimators of the parameters in GARCH processes. *Annals of Statistics*, 32, 633-655.
- Francq, C. and J-M. Zakoian (2004) Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli* 10, 605-637.
- Francq, C. and J-M. Zakoian (2007) Quasi-Maximum Likelihood Estimation in GARCH Processes when some coefficients are equal to zero. *Stochastic Processes and their Applications*, 117, 1265-1284.
- Francq, C. and J-M. Zakoian (2009) Optimal predictions of powers of conditionally heteroskedastic processes. Unpublished document.
- Hall, P. and Q. Yao (2003) Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica* 71, 285-317.

Patricia Reynaud-Bouret (CNRS-Nice)

**Adaptive estimation for Hawkes processes. Application to genome analysis,**  
joint work with *S. Schbath*.

This is a joint work with Sophie Schbath (INRA-Jouy en Josas). We provide a new practical method for the detection of either favored or avoided distances between genomic events along DNA sequences. These events are modeled by a Hawkes process. The biological problem is actually complex enough to need a non asymptotic penalized model selection approach. We provide a theoretical penalty that satisfies an oracle inequality even for quite complex families of models. The consecutive theoretical estimator is shown to be adaptive minimax for holderian functions with regularity in  $(1/2, 1]$ . Moreover we introduce an efficient strategy, named Islands, which is not classically used in model selection, but that happens to be particularly relevant to the biological question we want to answer. Since a multiplicative constant in the theoretical penalty is not computable in practice, we provide extensive simulations to find a data-driven penalty. The results obtained on real genomic data are coherent with biological knowledge and eventually refine them. There are questions that are still opened and I will mention them at the end of the talk.

**References:**

- Birgé, Lucien; Massart, Pascal Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields* 138 (2007), no. 1-2, 33–73.
- Daley, D. J.; Vere-Jones, D. An introduction to the theory of point processes. Vol. I et II. Second edition. *Probability and its Applications* (New York). Springer, New York, 2008.
- Gusto, Gaelle; Schbath, Sophie FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model. *Stat. Appl. Genet. Mol. Biol.* 4 (2005), Art. 24, 28 pp. (electronic).

Elodie Brunel (Montpellier II)

**Penalized contrast estimation of cumulative distribution functions in survival models.**

Most often in survival analysis, we do not observe directly the sample of interest  $(Y_1, \dots, Y_n)$  since lifetime data usually suffer from censorship. Different censoring mechanisms are encountered in applications. Some among them are random right-censoring, which is probably the most popular one, and to a lesser extent the interval censoring case 1, which often occurs in epidemiology or in actuarial sciences. For instance, such interval censoring arise in the study of infectious diseases, when the event of interest is an unobservable infection time. In this case, the only knowledge about the lifetime to infection  $Y$  is whether it has occurred before a medical examination time  $U$  or not. More formally, an observation consists of the pair  $(U, 1_{\{Y \leq U\}})$  where  $U$  is assumed to be independent of  $Y$ . Such sampling model may remind right-censored data where the observation is the pair  $(Y \wedge C, 1_{\{Y \leq C\}})$  and  $C$  stands for a censoring variable independent of  $Y$ . However, the estimation procedure in these censoring models is substantially different. Suppose that we are interested in the unknown cumulative distribution function (cdf)  $F(y) = \mathbb{P}(Y \leq y)$  of a lifetime  $Y$ . In the right-censoring model, the Kaplan-Meier estimator of the survival function is well studied and is known to be asymptotically normal (and uniformly consistent) at the rate  $\sqrt{n}$ . In the interval censoring, its counterpart is the nonparametric maximum likelihood estimator (NPMLE) introduced by Groeneboom and Wellner (1992) with convergence rate  $n^{1/3}$ . Another important feature in survival models is the presence of a covariable  $X$ . If so, the conditional cdf  $F(y|x) = \mathbb{P}(Y \leq y|X = x)$  should rather be considered.

We propose an overview of nonparametric strategies to deal with both interval and right-censoring and we address the problem of the adaptive estimation of the cdf and of the conditional cdf in presence of a real covariable  $X$ . We aim at proposing improvements in two directions. First, in the interval censoring case, our goal is to take into account the smoothness of the cdf to achieve the best convergence rate as possible, whereas in the right-censoring model, the same strategy would have a poor interest. On the other hand, if we are interested in the conditional cdf  $F(y|x)$ , a challenging nonparametric estimation problem is to find an estimator which is adaptive in the  $x$ -direction whereas in the  $y$ -direction no model selection is required. In both censoring settings, we consider regression-type estimators based on mean square contrasts. We also explain how to perform model selection by adequate penalization in the way developed by Barron et al. (1999). Then, our data-driven procedures lead to estimators that achieve automatically the optimal (minimax) rate.

**References:**

- Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, **113**, 301-413.
- Groeneboom, P. and Wellner, J.A. (1992). *Information bounds and nonparametric maximum likelihood estimation*. Boston, Birkhäuser Verlag.

**Erwan Le Pennec** (Paris VII)

**Maxisets for model selection,**

joint with *F. Autin, J.-M. Loubes* and *V. Rivoirard*.

In this talk, we characterize the statistical performance of model selection procedures in the classical Gaussian white noise model through the maxiset point of view. We present a general result for (almost) arbitrary models and specialize this result for wavelet based estimators.

We study penalized estimator in the classical Gaussian white noise model

$$dY_{n,t} = s(t)dt + \frac{1}{\sqrt{n}}dW_t, \quad t \in \mathcal{D},$$

where  $\mathcal{D} \subset \mathbb{R}$ ,  $s$  is the unknown function,  $W$  is the Brownian motion and  $n \in \mathbb{N}^* = \{1, 2, \dots\}$ . We consider a collection of models  $\mathcal{M}_n$  spanned by some functions  $\phi_i$  of a dictionary  $\Phi$ . For each model  $m$ , we define the estimate  $\hat{s}_m$  that minimizes the quadratic empirical criterion

$$\gamma_n(u) = -2 \int u dY_{n,t} + \|u\|^2 \quad \text{over } u \in m.$$

Now, the issue is the selection of the best model  $\hat{m}$  from the data which gives rise to the *model selection estimator*  $\hat{s}_{\hat{m}}$ . Following a long tradition, we propose to select the model  $\hat{m}$  with the following penalized criterion

$$\hat{m} = \arg \min_{m \in \mathcal{M}_n} \left\{ \gamma_n(\hat{s}_m) + \frac{\lambda_n}{n} D_m \right\}$$

where  $D_m$  is the dimension of  $m$  and  $\lambda_n$  is a “large enough” constant that may depends on  $n$ .

Characterizing the performance of such an estimator is an important issue. Those penalized estimator have been already extensively studied. The quadratic risk  $E[\|s - \hat{s}_{\hat{m}}\|^2]$  has been shown to be smaller up to a known factor than the deterministic quantity

$$Q(s, n) = \inf_{m \in \mathcal{M}_n} \left\{ \|s_m - s\|^2 + \frac{\lambda_n}{n} D_m \right\} .$$

This has been used to exhibit, for specific dictionary, function classes on which those estimators are minimax. Following the maxiset approach, we characterize the sets of functions estimated at a given rate. We establish a general result that does not require to specify the dictionary used. More precisely, we establish an equivalence between the statistical performance of  $\hat{s}_{\hat{m}}$  and the approximation properties of the model collections  $\mathcal{M}_n$ . For

$$\rho_{n,\alpha} = \left( \frac{\lambda_n}{n} \right)^{\frac{\alpha}{1+2\alpha}}$$

for any  $\alpha > 0$ , we prove that, for a given function  $s$ , the quadratic risk  $E[\|s - \hat{s}_{\hat{m}}\|^2]$  decays at the rate  $\rho_{n,\alpha}^2$  if and only if the deterministic quantity  $Q(s, n)$  decays at the rate  $\rho_{n,\alpha}^2$  as well. This result holds with mild assumptions on  $\lambda_n$  and under an embedding assumption on the model collections ( $\mathcal{M}_n \subset \mathcal{M}_{n+1}$ ).

We apply then these results to some wavelet estimators (a linear strategy, a threshold strategy and a mixed strategy) and characterize exactly their maxisets.

### References:

- Cohen A., DeVore R.A., Kerkyacharian, G. and Picard, D. *Maximal spaces with given rate of convergence for thresholding algorithms*. Appl. Comput. Harmon. Anal. **11**, no. 2, 167-191, 2001.
- Daubechies, I. *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.
- DeVore, R.A. and Lorentz, G.G. *Constructive approximation*. Springer-Verlag, Berlin, 1993.
- Kerkyacharian, G. and Picard, D. *Thresholding algorithms, maxisets and well-concentrated bases*. Test 9, no. 2, 283-344, 2000.
- Massart, P. *Concentration inequalities and model selection* Lectures on probability theory and statistics (Saint-Flour, 2003), Lecture Notes in Math., 1896, Springer, Berlin, 2007.

Christophe Pouet (Aix-Marseille I)

**Test on components of densities mixture,**  
joint work with *F. Autin*.

This talk deals with statistical tests on components of densities mixture. We propose to test whether the densities of two independent samples of independent random variables  $Y_1, \dots, Y_n$  and  $Z_1, \dots, Z_n$  come from the same mixture of  $M$  components or not. This testing problem was recently considered by Butucea and Tribouley (2006). More precisely let  $Y_1, \dots, Y_n$  be a sample of independent random variables with unknown marginal densities:

$$f_i(\cdot) = \sum_{u=1}^M \omega_u(i) p_u(\cdot), \quad 1 \leq i \leq n.$$

and let  $Z_1, \dots, Z_n$  be an other sample of independent random variables with unknown marginal densities

$$g_i(\cdot) = \sum_{u=1}^M \sigma_u(i) q_u(\cdot), \quad 1 \leq i \leq n,$$

which is independent from the sample  $Y_1, \dots, Y_n$ .

We assume that the weights  $(\omega_u(i), 1 \leq u \leq M, 1 \leq i \leq n)$  and  $(\sigma_u(i), 1 \leq u \leq M, 1 \leq i \leq n)$  are known and satisfy some conditions. We consider the same set-up as Maiboroda (2000) and Pokhyl'ko, D. (2005).

We provide test procedures which are optimal according to the minimax setting and compute the minimax rate of testing. Moreover we extend our results to the adaptive case. We provide an adaptive test procedure and establish the lower bound. One of the most interesting point in this problem is the role played by the weights. We will discuss it and show simulations to illustrate their influence in practice. We will also discuss the case of random weights and give some hints for further developments.

**References:**

- Butucea, C., and Tribouley, K. (2006). Nonparametric homogeneity tests. *Journal of Statist. Plan. and Inf.*, vol. **136**, 597-639.
- Maiboroda, R.E. (2000). A homogeneity criterion for mixtures with varying concentrations. *Ukrainian Math. J.*, vol. **52** (8), 1256-1263.
- Maiboroda, R.E. (2000). An asymptotically effective estimate for a distribution from a sample with a varying mixture. *Theory Probab. Math. Statist.*, vol. **61**, 121-130.
- Pokhyl'ko, D. (2005). Wavelet estimators of a density constructed from observations of a mixture. *Theor. Prob. and Math. Statist.* vol. **70**, 135-145.