Dependence in Statistics with applications

My work makes use of the dependence of stochastic processes, for which I am one of the world-leaders, for statistical inference. Randomness is an easy way to model phenomena for which it looks impossible to propose a deterministic model; think of the stock-exchanges, the quality control in the industry, meteorological forecasts or biological problems in computational genomics. Dependence between data clearly occurs in many time varying problems as well as in questions where geographic questions arise. Dependence definitely interacts with the random behaviors. The project is aimed at developing real statistical applications of dependence. The Short and Long Range Dependence (resp. SRD and LRD) will be considered in both the discrete and continuous-time cases, and also in the case of random fields (geographic dependence). Note that asymptotic results in case of SRD look similar to independent situations up to some constants. At the contrary LRD yields new normalizations. Valuable models should be introduced in order to fit the properties of current data: integer valued, continuous time or point process data will be considered. The identification of the models is an essential question; tests of goodness-of-fit are also needed to check their validity. The level of reliability of the procedures may be calculated from empirical methods. The previous problems give a special importance to the resampling techniques. This allows quantifying the quality of the calibration procedures. Applications will be developed from the methodological viewpoint, the computational implementation should also be considered by actualization of the standard statistical packages. Adapted massive calculations techniques will be developed. Applications of dependence will concern finance and insurance, demography, biology and several questions in ecology.

My work is centred on the dependence of stochastic processes. In 1956 Pr. Murray Rosenblatt introduced a strong mixing condition extending on those from ergodic theory. A huge success of such notions is due to its simplicity and to the powerful range of mathematical applications of those notions. The mixing setting is widely used in econometry. I also worked during years with those notions and my monograph on mixing is among the most widely cited in this domain (789 citations on Google Scholar).
I also introduced some very widely used tools of non parametric statistics like wavelets considered in a short note at Comptes Rendus Academy of Sciences 1988. Moreover other questions related to super-resolution or continuous time and functional estimation techniques were solved (see the short bibliography). I am cited 412 times by 311 authors on mathscinet and almost 2000 times on Google Scholar.

The asymptotic properties of weak or strong dependence are central in my research work. I am very involved in the problem of modelling times series with specific properties.

- I introduced several nonlinear models exhibiting distributional long range dependence and the monograph [2] is a main reference for long range dependence (LRD).
- For the weakly dependent case (WD), my monograph [1] is among the most cited reference for mixing and a recent work (see the monograph [5]) proposes a real alternative to mixing to measure weak dependence.

My aim is to go further on in the study of the statistical problems linked with dependence during this program of research. Applications should be definitely addressed up to their complete realization.

Extended Synopsis of the project

The present proposal is focused on the dependence of stochastic processes. Dependence of the time series is an essential feature of real applications. Very specific problems occur with this particular and natural property. Dependence under consideration concerns together time series, continuous time processes, random fields, or space-time processes. This means that random observations are respectively observed according discrete epochs like years or seconds, or continuously; and respectively observations are geographic (according to some graph relations) or even combining geographic and time indices.

The first and really important problem is the way to define the dependence conditions. In 1956 Pr. Murray Rosenblatt introduced a strong mixing condition extending on those from ergodic theory. The concept appears in ergodic theory and strong mixing techniques were introduced in order to takle statistical applications. A huge success of such notions is due to its simplicity and to the powerful range of mathematical applications of those notions. This *mixing* setting is widely used in Econometry. I also worked during years with those notions and Doukhan (1994) is among the most widely cited publications on mixing (789 citations on Google Scholar).

A very simple example of an autoregressive model with discrete inputs was exhibited in 1984 by Donald Andrews. I thus introduced a new and very simple notion of weak dependence in 1999 with Sana Louhichi. This frame is definitely a fruitful one as stressed in our collective volume Dedecker *et al.* (2007). I also introduced several nonlinear models exhibiting distributional long range dependence and the monograph Doukhan *et al*. (2003) is a main reference for long range dependence.

Limit theorems in probability theory under mixing and other dependence conditions are thus studied for the sake of applications. Problems related to the non-parametric estimation as well as fitting models to real data in statistics are among those applications. These and some other applications in finance and stochastic algorithms are already discussed in my preceding papers but there is still a number of open questions and unsolved problems concerning statistical inference and forecasting problems. The program in this proposition is to develop such applications for the next years.

Among the following important items the proposal is aimed at finding explicit solutions for applications for which a statistical dependence is relevant. This means that priority will be given to those applications without avoiding deep theoretical questions. More precisely the selected applications are essentially those for which improvements of the theory are really needed. Applications to Ecology (hydrology, disposition of the trees in a forest, trajectories of shoals of fishes and global warming), in biology (through sparsity techniques and applications to *DNA* codes), and to finance (through the estimation of extremes tails and the behaviour of high frequency data) will need much modelling efforts as well as many improvement of the statistical theory.

1 Modelling integer valued times series

The first important example from the application viewpoint is that of integer valued time series. Let us give some examples: in quality control when we observe the number of defective items found in successive samples taken from a production line; in epidemiology in identifying the numbers of cases of important diseases like *AIDS* in a given area in successive months; in road security surveillance, the number of deadly car accidents in successive weeks; in finance, the presence or absence of trading in a particular share on consecutive trading days. Monitoring of a specific disease is an important task in social and preventive medicine. Some integer-valued models have been used to represent times series like the fortnight number of campylobacter and meningitis cases in a given area. See for example the models proposed by Ferland *et al.* (2006)

or Doukhan *et al.* (2006). There exist regression models for time series of counts, see (Kedem and Fokanios, 2002, Chapter 4). Another point of view for defining discrete values processes is presented in Garcia *et al.* (2009) where applications to linguistic are also given. The models used here are infinite memory discrete time series which fit some weak dependence conditions.

2 Continuous time processes and random fields

Modelling *continuous times processes* with weak dependence properties is a real challenge of a current main interest. A special attention should be given to Lévy driven diffusion processes. Non-Markov models such as equations with delay driven with Lévy processes will also be proved to exist. They are extensions of discrete time series denoted *LARCH* models in Dedecker *et al.* (2007). A project submitted to the French Agency for Research is to detect the trajectories of shoal of fishes; such trajectories are modelled as discretely observed vector valued continuous time processes. Prediction is also a natural feature of this problem. Either short-range or long-range dependence may be exhibited. Hence this questions enlightens several problems: we first need to propose continuous times models with weak dependence; the question of defining vector valued long range dependence is the second important topic of the question. Models with continuous time of diffusions with jumps are adapted to deal with risk management. Inference and prediction of such data are clearly essential for financial purpose.

Prediction of space-time data. Random fields modelling is another problem of interest beyond Dobrushin type conditional models, in Doukhan and Truquet (2007) we present alternative models, solutions of intrinsic equations $X_t = F((X_{t-s})_{s\neq 0}, \epsilon_t)$. Coherent models with space-time indices will be introduced; their coherence should be of a physical origin. Applications to meteorology where many such models have been introduced will be a main subject. Namely the problem of global warming will be considered according to the ideas introduced for the simpler time dependent case.

3 Dependence and point processes

An item of research related with extremes is to investigate more accurately the dependence properties of point processes. Such models appear naturally, they denote a realization of points thrown on a plane according to distribution without finite limit points; the number of points is thus denumerable and observation is given in a bounded zone so that it is a finite random set. Besides the homogeneous Poisson process which correspond to an ideal independent situation, many types of Markov or compound point processes need to be considered (see Daley and Vere-Jones, 2003). Notions of dependence adapted to point processes are also needed. Applications to ecology problems may be found in Lang and Marcon (2010). General dependence conditions of point processes need to be introduced. Another project is concerned by a problem of ecology; the location of trees in a forest needs a description to be proposed as clustering models for forests. Intra and inter species are of a specific interest for ecological modelling. Such questions are addressed by the National Agency of Waters and Forests **(ENGREF).**

4 Long range dependence (LRD)

LRD phenomena were exhibited in Rosenblatt (1962). A main interest was stressed for the corresponding limit behaviours because of their fractal behaviour. Such processes were first used for modelling the fjords in Norway and Benoit Mandelbrot (1979) also used the first example of fractional Brownian motion to this aim. In this case the Hurst exponent determines also the regularity of trajectories. Hurst (1951) exponent was introduced in an hydrology problem in order to optimize the size of a reservoir. Analysis of stock exchange values also fits such fractal models. LRD is often assimilated with a short list of simple models, which are often not completely satisfactory. More complicate models are functions of linear processes and a nice treatment of their properties follows from Ho & Hsing (1997), the case of models $X_n = G(Y_n, \epsilon_n, \epsilon_{n-1}, \dots)$ excited by a linear process is processed in a join work with Lang and Surgailis (2010). Such models are proved to follow a more specific type of asymptotic behaviour

that the linear or Gaussian case. For those results, the existence of moments with order more than 2 has to be assumed. Heavy tailed *LRD* appears in the field of traffic in telecommunications (empirical studies prove that moments of order 2 are not finite), see Willinger *et al.* (1996); however traffic is a complex field where non-stationary, heavy-tails and external phenomena play competitive roles, moreover new questions such as streaming are now of a main importance and they seem hard to model here. Samorodnitsky and Taqqu (1994) propose a deep survey of this problem. An important paper for limit theory of partial sums in this case is Mikosch *et al.* (2002). Limit theory is considered for empirical cdf of linear processes in Surgailis (2002), and for some random *AR*-processes in Leipus *et al.* (2006); we also refer a very nice recent work by Bartkiewicz *et al.* (2009) but anyway few is known under general conditions even for this simple partial sums process. We aim at considering a more systematic class of such models.

5 Non stationarity

Non-stationarity may be modelled in different ways. A first way to model non-stationary times series is to write for some stationary times series (ξ_k) and two regular functions a, and b on the time interval $[0,1]$

$$X_k = a(j/n) + b(j/n)\xi_k \quad for \quad 1 \le k \le n$$

The estimation of those coefficients is based on local partial sums processes and other kernel-type methods and isotonic regression may also be used. Interesting situations appear when the error process ξ_k in this model is long range dependent. This often occurs when considering financial or climatology time series. For instance the annual series of winter means of the NAO index (North Atlantic Oscillation index) exhibits long range dependence (see Dedecker *et al.* 2009). Horvath also considered estimation for non stationary times series. More accurate models of non stationary processes will be developed together with asymptotic theory for extremes of such models. Moreover alternatives between non stationarity is an interesting other practical issue. Global warming is a main question and considering space-time dependence is thus natural way to visualize its progression and to help its comprehension.

6 A program of statistical data analysis

An important step is indeed to decide if data considered are *stationary* or not (and also to discuss the counter-hypothesis, weak dependence or long range dependence); then a simple transformation, maybe an affine one, will transform the data into a stationary time series. Dahlhaus (2009) proposed interesting views to this question extended in further work on progress by Neumann. The idea is to prove second order stationarity by considering the difference of an integrated local spectrum and the standard periodogram; null hypothesis leads to a statistical decision procedure. *Changes of regime* may also be analyzed through classical techniques or through Surgailis *et al.* (2008)'s *IR* statistics. *Testing LRD* versus *SRD* will be exhibited by using the very distinct behaviours of the empirical process under those two conditions.

Fitting dependent models: after this, one should select a suitable model (among all those previously defined) for the proposed data. *Resampling* and tests for goodness of fit test are useful to decide of a suitable model. *E.g.* a dependent analogue of Kolmogorov-Smirnov test based on the empirical cumulative function will be developed; in this dependent setting a main problem is that, contrarily to the independent case, the sup-bound of the limit process Z of the empirical central limit theorem admits a distribution which relies on the whole dependence structure of (X_t). Statistical validation of the model is provided through previously introduced questions (in particular *Whittle* estimator or QMLE will be investigated for the above mentioned models). In the continuous time case we will also need to validate *discretization* procedures. *Model selection techniques* classically based on concentration inequalities will be developed. Non parametric tests for goodness-of-fit will also be considered; subsampling techniques are essential to estimate the quantiles of this distribution. *Prediction* will be processed through both regression estimation and quantile regression estimation.

7 Statistical applications

Limit asymptotic theory for various statistics of interest is often a difficult question, both from the theoretical and from the computational viewpoint. Bickel and Bühlmann (1999) (which is a real counterpart of Doukhan & Louhichi, 1999) provide tools in this direction by *bootstrapping time series*. Such methods seem useful for extremes, records, unit-root problems, or *U*-statistics; see Neumann and Paparoditis (2008). An alternative to bootstrap are the methods of *subsampling,* see Bertail *et al.* (2004). A smoothed version of subsampling adapted to work with weak dependence is presented in Doukhan *et al.* (2010); in this paper the example of subsampling for extremes is essential.

8 Stochastic Algorithms

Algorithms with dependent inputs were already discussed by Brandière and Doukhan (2004). An intensive use of such algorithms is necessitated for portfolio optimization. The topic *Randomized algorithms* has roots in computer science and applies to computer algorithms that, in addition to input, take a source of random numbers and make random choices during execution. For many problems, a randomized algorithm is the simplest or the fastest. Some randomized algorithms and probabilistic analysis for them is discussed by Klesov (2010).

The problem of *sparsity* has received a lot of attention in the last few years. There is no theoretical result on the *LASSO* (see Tibshirani, 1996) in the case where the data are dependent. However, these results can be easily obtained: the only use of the fact that the data are independent is in Bernstein's inequality. Using Bernstein's inequality for dependent data (see Doukhan and Neumann, 2006), this looks possible to generalize the results of Bickel *et al.* (2007) to the case of dependent data. However, it would be even more interesting and more difficult to study also the *LASSO* in the case of an auto-regression. A study of their asymptotic and non-asymptotic property is thus needed.

9 Finance and extremes of stationary sequences

Value-at-Risk *(VaR)* has been introduced from 1980s, in particular from 1987 market crash; it plays a central role in financial risk management and control. This risk measure is widely used not only to quantify the risk of financial portfolio losses but also to regulate the whole financial industry. In 1999, the Basel II Accord promoted the use of *VaR*. Most of the standard literature about extreme events assume independence (see *e.g.* Embrechts *et al.*, 1997). But, as illustrated by the recent financial crisis, such assumption is not reasonable. It is the reason why there exists an urgent need for introducing dependence to provide better risk control.

Extremes occur clearly in for financial or insurance purpose but many other type of applications like geophysics. Earthquakes or tsunami are at the origin of the theory of extremes; they will be investigated through questions arising from the petroleum industry. The exact distribution of the supremum of specific random processes and fields is often a difficult question. Asymptotic theory of extreme values of a time series is thus useful. Additional assumptions on the regularity of tails of the underlying distributions and/or on the mixing allow one to evaluate precisely the distribution of a large sample of a times series. Referring to Leadbetter, Lingren and Rootzen (1983) the asymptotic behavior of times series is « the same as » for independent and identically distributed sequences under two assumptions **D'** and **D**; the first one is an anticluster condition while the second one is a « tail-mixing » assumption. Under an additional tail regularity assumption an affine transform of the sequence of maxima converges in distribution to a distribution depending on a Pareto parameter α and an extremal index θ. Anyway various models do not satisfy those conditions (see Hall *et al.* 2009), in particular for real data, and even in case all the conditions are well known complicated normalizations make the estimation of a quantile of extreme values quite numerically unstable. Moreover without any additional information one may address the estimation of the quantiles of such extremes (usually called as Values at Risk in

Finance). Subsampling is a first useful tool for this topic but in Finance, sample sizes are often not large enough for such techniques (see Bertail *et al.*2004). A more adapted way to estimate such *VaR*'s is to use resampling for parametric times series models such as in Sixiang Cai's doctoral work. To this aim various steps are due. First a deep estimation work is essential. Second one estimates the residuals in order to run the resampled version of a times series model. Third one works out statistics of interest; in our case this is either extremes or higher order quantiles. The point process of exceedances is of a main interest. Indeed the interesting value is certainly a high order quantile, with level tending to 1 with the sample size, more that the extreme of a times series. A nice review book on extremes is the collective one edited by Andersen *et al*. (2009).

Bibliography

Andersen, T.G., Davis, R.A., Kreiß, J-P., Mikosch, T. *editors* (2009) *Handbook of Financial Time Series*, Springer Berlin Heidelberg.
Andrews, D. (1984) *Non strong mixing autoregressive processes.* J. of Appl. Probab **21**, 930-934.
Bartkiewicz, K,. Jakubowski, A., Mikosch, T., Wintenberger, O. (2009) *Infinite variance stable limits for sums of dependent variables.* Preprint.
Beran, J. (1994) *Statistics for long-memory processes.* Monographs on Statistics and Applied Probability **61**. Chapman and Hall, New York.
Bertail, P., Haefke, C., Politis, D.N., White, W. (2004) *Subsampling the distribution of diverging statistics with applications to finance* Journal of Econometrics **120**, 295-326.
Bickel, P., Ritov, V., Tsybakov, A. (2007) *Simultaneous analysis of LASSO and Dantzig selector*. Annals of Statistics **37**, 1705-1732.
Bickel, P., Bühlmann, P. (1999) *A new mixing notion and functional central limit theorems for a sieve bootstrap in time series.* Bernoulli **5**-3, 413-446.
Brockwell, P.J., Davis, R.A. (2002) *Introduction to time series and forecasting.* Springer Texts in Statistics. New York. Springer-Verlag, second edition.
Chernick, M. (1981) *A limit theorem for the maximum of autoregressive processes with uniform marginal distribution.* Annals of Probability **9**, 145-149.
Dahlhaus, R. (1997) *Fitting time series models to nonstationary processes.* Ann. Statist. **25**, 1-37.
Dahlhaus, R. (2009) *Local inference for locally stationary time series based on the empirical spectral measure.* Journal of Econometrics **151**, 101–112.
Daley, D.J., Vere-Jones, D. (2003) *An Introduction to the Theory of Point Processes **1**, Elementary Theory and Methods.* Springer-Verlag, New York, 2nd Edition.
Davis, R.A., Mikosch, T. (2009) *The extremogram: A correlogram for extreme events.* Bernoulli **15**-4, 977-1009.
Dedecker, J., Doukhan, P., Lang, G., León, J.R., Louhichi, S., Prieur, C. (2007) *Weak dependence: models, theory and applications.* Lecture Notes in Statistics **190**, Springer-Verlag, New-York.
Dedecker, J., Merlevède, F., Peligrad, M. (2009) *Invariance principles for linear processes with application to isotonic regression.* Preprint.
Doukhan, P. (1994) *Mixing: Properties and Examples.* Lecture Notes in Statistics **85**, Springer-Verlag, New-York.
Doukhan, P., Lang, G., Surgailis, D. (2010) *A class of Bernoulli shifts with long memory asymptotics of the partial sums process.* Submitted.
Doukhan, P., Latour, A., Oraichi, D. (2006) *A simple integer-valued bilinear times series models.* Advances in Applied Probability **38**, 1-20.
Doukhan, P., León, J.R., Nicolle, J.L. (1988) *Metodologia para evaluar la sismicidad cuando la base de datos es incompleta*. Revista Tecnica INTEVEP **8**.1, 13-22.
Doukhan, P., Louhichi, S. (1999) *A new weak dependence condition and applications to moment inequalities.* Stochastic Processes and their Applications **84**, 313-342.
Doukhan, P., Neumann, M. (2007) *A Bernstein type inequality for times series.* Stochastic Processes and their Applications **117**-7, 878-903.
Doukhan, P., Oppenheim, G., Taqqu, M., *eds.* (2003) *Theory and Applications of Long-range Dependence.* Birkhäuser, Boston.
Doukhan, P., Prohl, S. Robert, C.Y. (2010) *Subsampling weakly dependent times series and application to extremes.* Submitted.
Doukhan, P., Truquet, L. (2007) *A fixed point approach to model random fields.* Alea Latino-American Journal of Probability and Statistics **2**, 111-132.
Embrechts, P., Klüppelberg, C ., Mikosch, T. (1997) *Modelling Extremal Events*. Springer, Berlin.

Ferland, R., Latour, A., Oraichi, D. (2006) *Integer-valued GARCH process.* Journal of Time Series Analysis **27**, 923-942.
Garcia, N., Galves, A., Prieur, C. (2009) *Perfect simulation of the **d**-minimal coupling between ordered pairs of binary chains of infinite order.* Submitted.
Ho, H.-C., Hsing, T. (1997) *Limit Theorems for Functionals of moving averages.* Annals of Probability **25**-4, 1636-1669.
Kedem, B., Fokanios, K. (2002) *Regression Models for Time Series Analysis.* Wiley series in probability and statistics. Hoboken, New Jersey: Wiley-Interscience.
Klesov, O.I. (2010) *Randomized algorithms and probabilistic analysis*. Kiev, TBiMC (in Russian).
Lang, G., Marcon, E. (2010) *Ripley statistic as a test for the Poisson point process hypothesis*. Submitted to the Annals of Statistics.
Neumann, M., Paparoditis, E. (2008) *Goodness-of-fit tests for Markovian time series models*. Bernoulli **14-**1, 14-46.
Hall, A., Scotto, M. G., Cruz, J. (2009) *Extremes of integer-valued moving average sequences.* Test.
Rosenblatt, M. (1956) *A central limit Theorem and a strong mixing condition.* Proceedings of the National Academy of Sciences USA **42**, 43-47.
Rosenblatt, M. (1961) *Independence and dependence,* in Proceeding **4**th. Berkeley Symposium Mathematical. Statistics and Probability 411-443. Berkeley University Press.
Rosenblatt, M. (1991) *Stochastic curve estimation.* NSF-CBMS Regional Conference Series in Probability and Statistics **3**.
Surgailis, D., Teyssière, G., Vaičiulis, M. (2008) *The increment ratio statistic.* Journal of Multivariate Analysis **99**-3, 510-541.
Tibshirani, R. (1996) *Regression shrinkage and selection via the LASSO.* Journal of the Royal Statistical Society Ser. B, **58**-1, 267-288.
Willinger, W., Taqqu, M.S., Erramilli, A. (1996) *A bibliographical guide to self-similar traffic and performance modeling for modern high- speed networks, Stochastic Networks: Theory and Applications*, F.P. Kelly, S. Zachary and I. Ziedins, *editors*, 339-366, Clarendon Press, Oxford.