Actes des Deuxièmes Journées en Statistique Fonctionnelle et Opératorielle

12-13 Juin 2003, Toulouse

Recueillis par
S. Viguier-Pla, A. Boudou, H. Cardot, F. Ferraty
Y. Romain, P. Sarda, P. Vieu, et A. Yousfate

DEUXIÈMES JOURNÉES DE STATISTIQUE FONCTIONNELLE ET OPÉRATORIELLE
12-13 Juin 2003, TOULOUSE

Présentation des journées

Sylvie VIGUIER-PLA*, Alain BOUDOU, Hervé CARDOT, Frédéric FERRATY, Yves ROMAIN, Pascal SARDA, Philippe VIEU, et Abderrahmane YOUSFATE

* Présidente du Comité Scientifique et d'Organisation
Laboratoire Statistique et Probabilités, Université Paul Sabatier
31062 Toulouse Cedex
e-mail : viguier@cict.fr

Le groupe de travail STAPH en Statistique Fonctionnelle et Opératorielle en est à sa quatrième année d'existence (voir [1], [2], [3], [5]), et les retombées sont nombreuses et prometteuses. Il est par exemple indéniable que les aspects fonctionnels sont de plus en plus présents en Statistique : que ce soit par exemple au travers des techniques et/ou modèles non-paramétriques, au travers de la vision opératorielle de la Statistique, au travers des techniques et/ou modèles pour l'analyse et/ou le traitement de courbes ...

Ainsi, et au vu du succès de la première édition (voir [4]), nous avons décidé de terminer cette année universitaire en organisant pour la seconde année consécutive des Journées d'échanges scientifiques portant sur "les divers aspects fonctionnels" de la Statistique.

Nous avons délibérément décidé de placer ces journées sous le signe de l'ouverture scientifique, en accordant par exemple une place prépondérante aux intervenants extérieurs à Toulouse. Nous avons aussi réservé une grande place aux jeunes chercheurs qui constituent la plupart des intervenants de ces journées.

Notre démarche se situe toujours, comme l'année dernière, à la fois dans le cadre d'une collaboration entre de nombreuses équipes toulousaines de statisticiens (Univ. Paul Sabatier, Univ. Toulouse le Mirail, INRA) et dans celui d'une coopération avec nos collègues statisticiens de l'Université de Sidi Bel Abbès en Algérie. Tous ces organismes de recherche, par leur participation scientifique active, contribuent pleinement à la réussite de notre initiative et nous souhaitons les en remercier vivement.

Il va de soi qu'une telle manifestation n'aurait pas pu avoir lieu sans soutien financier, et nous souhaitons remercier vivement les efforts en ce sens consentis

par l'Université Paul Sabatier, l'INRA, l'Université de Sidi bel Abbes, le Laboratoire de Statistique et Probabilités de Toulouse ainsi que le support financier de la Société Leybold Didactic GmbH[1]. La participation de Françoise Michel aux diverses tâches administratives et à l'organisation des pauses café a grandement contribué au bon déroulement et à l'ambiance détendue de ces journées, et nous souhaitons la remercier vivement.

Pour terminer, nous souhaitons rappeler l'existence de notre page web (voir [6]), sur laquelle les intéressés peuvent accéder à l'ensemble de nos activités. Enfin, signalons que pour l'année à venir notre groupe de travail est impliqué dans l'organisation de deux sessions Statistique Fonctionnelle et Opératorielle lors du Congrès Canada-France des sciences mathématiques en Juillet 2004 (voir [7]) et lors du congrès COMPSTAT 2004 (voir [8]).

Merci encore à tous les participants, conférenciers ou auditeurs, ...

Références

1 Staph, (2001a). Statistique Fonctionnelle I : Groupe de Travail STAPH, Résumés des Exposés, Années 1999-2000. *Publication du laboratoire de Statistique et Probabilités*, **LSP-2001-05**, Toulouse.

2 Staph, (2001b). Statistique Fonctionnelle II : Groupe de Travail STAPH, Résumés des Exposés, Années 2000-2001. *Publication du laboratoire de Statistique et Probabilités*, **LSP-2001-07**, Toulouse.

3 Staph, (2002a). Statistique Fonctionnelle III : Groupe de Travail STAPH, Résumés des Exposés, Années 2001-2002. *Publication du laboratoire de Statistique et Probabilités*, **LSP-2002-12**, Toulouse.

4 Staph, (2002b). Actes des Premières journées de Statistique Fonctionnelle, Toulouse Juin 2002, *Publication du laboratoire de Statistique et Probabilités*, **LSP-2002-09**, Toulouse.

5 Staph, (2003). Statistique Fonctionnelle IV : Groupe de Travail STAPH, Résumés des Exposés, Années 2002-2003. *Publication du laboratoire de Statistique et Probabilités*, Toulouse, en préparation.

6 *http ://www.lsp.ups-tlse2.fr/Fp/Ferraty/staph.html*

7 *http ://smc.math.ca/Reunions/Toulouse2004*

8 *http ://compstat2004.cuni.cz*

[1] le bilan financier détaillé est accessible sur notre page web ([6])

Using density level sets for nonparametric control charts

Amparo BAILLO *

* Adresse pour correspondance :
Departamento de Estadística y Econometría
Universidad Carlos III de Madrid,
28903 Getafe (Madrid). Espagne.
e-mail : abaillo@est-econ.uc3m.es

Introduction

A large part of the literature on FDA has focused on functions of a single argument, $x(t)$, like pinch force, height or temperature recorded over time (see Ramsay and Silverman 1997 and references therein). In these settings the i-th. individual of the sample is characterised by a function $x_i(t)$. Nevertheless there are often situations where data are organized by space. In particular, there are powerful imaging techniques, such as magnetic resonance or tomography that allow observation of, for example, changes in brain shape or size. Occasionally these techniques reveal substantial shape differences in the brain of patients suffering from different illnesses.

Resonance and tomography can be placed in the general framework of stereology, the reconstruction of a body from lower-dimensional sampled information (see, e.g. Stoyan, Kendall and Mecke 1995). For instance, in the three-dimensional case one could think of estimating a body from random sections of dimension two (hyperplanes), one (straight lines) or zero (points). In this context and also in image analysis, each individual from the sample is represented by a function recorded over space. The FDA study of this function in each of the different groups would maybe lead to a discriminant rule for correctly diagnosing a patient or for adequately classifying an image.

Another, somewhat simpler, target in the general framework of image analysis is set estimation, which deals with the statistical problem of estimating an unknown (generally compact) set $S \subset \mathbb{R}^d$ from a sample of identically distributed points X_1, X_2, ..., X_n, randomly selected in S. In the language of probability theory this amounts to the estimation of the support of the common underlying distribution of the X_i. A closely related more general problem is that of estimating level sets of type $\{f > c\}$, where f stands for the density of X_i and $c > 0$

is a given constant. Or we could also consider estimating the boundary of a set. For example, Rudemo and Stryhn (1994) considered the problem of classifying a collection of leaves in different plant species according to their shape.

The works by Rényi and Sulanke (1963) and Geffroy (1964) are often cited as pioneering references in set estimation. Most results on this subject deal with the case where S is assumed to be convex. Under this assumption the natural estimate of S is the convex hull of the sample (see Efron 1965, Moore 1984, Schneider 1988, Dümbgen and Walther 1996). In the general case (non-convex S) there is no unique natural estimator of S. Maybe the simplest alternative is

$$\hat{S}_n = \bigcup_{i=1}^{n} B(X_i, \epsilon_n), \tag{1}$$

where $B(x, a)$ denotes the closed ball centered at x with radius a, and ϵ_n is a sequence of smoothing parameters which should tend to zero slowly enough in order to achieve consistency. Some properties of this estimator can be found in Chevalier (1976), Devroye and Wise (1980), Grenander (1981), Cuevas (1990) and Korostelev and Tsybakov (1993). An interesting application of (1) to satellite image analysis is given by Bertholet, Rasson and Lissoir (1998).

The mathematical methodology used in the non-convex case is closer to the spirit of nonparametric functional estimation. This is not surprising since the problem of set estimation can be reduced to that of estimating an indicator function. Furthermore, in estimator (1), the sequence ϵ_n plays a role analogous to that of the bandwidths in kernel estimation (see, e.g., Simonoff 1996). One of the main differences is, of course, that suitable metrics for sets, rather than functional metrics, are required here. We will use the measure of symmetric difference

$$d_\mu(S, T) = \mu(S \Delta T),$$

where μ is a σ-finite measure on $\mathbb{R}^d$ and Δ denotes the symmetric difference between sets, $S \Delta T = (S \setminus T) \cup (T \setminus S)$. This pseudodistance is reminiscent of the L_1 metric in density estimation.

Some methods of nonparametric functional estimation arise as very useful tools in set estimation problems : under mild regularity conditions on the underlying density f, the support S is essentially the set $\{f > 0\}$. A plug-in estimator would be $\{f_n > 0\}$ (where f_n is a nonparametric density estimator of f), although estimators of the type $\{f_n > c_n\}$ with $c_n \downarrow 0$ actually provide a richer and more flexible alternative (see Cuevas and Fraiman 1997 and Molchanov 1998).

Density level sets play a main role in cluster analysis. Given a density f, the population c-clusters are defined, following Hartigan (1975), as the connected components of the level set $\{f > c\}$. Hence level set estimation appears as a natural intermediate step. We will be interested in the plug-in estimator

$$\tilde{S}_n = \{f_n > c\}, \tag{2}$$

where f_n denotes a kernel estimator $f_n(x) = (nh^d)^{-1} \sum_{i=1}^{n} K((x - X_i)/h)$, with kernel K and bandwidth $h = h_n$, such that K is a density, $h_n \to 0$ and $nh_n^d \to \infty$.

In the nonparametric setting, estimation of level sets was considered, from a theoretical viewpoint, in Polonik (1995), who proposed the so-called excess mass approach, and in Tsybakov (1997). Some nonparametric, computationally feasible, clustering techniques can be found in Walther (1997) and Cuevas, Febrero and Fraiman (1999).

Level sets have also an obvious interpretation in terms of confidence sets. In a parametric framework, Davies and Gather (1993) define an α-outlier Z as the observation such that $Z \in \{f \leq c\}$, where

$$\int_{\{f \leq c\}} f = \alpha, \tag{3}$$

for a given $\alpha \in (0, 1)$. In general f is unknown and we are interested in the straightforward extension that (3) has in the case of the plug-in estimator (2), when the constant c is substituted by a random value c_n satisfying

$$\int_{\{f_n \leq c_n\}} f_n = \alpha \quad \text{a.s.} \tag{4}$$

Then an observation Z would be declared to be an α-outlier if $Z \in \{f_n \leq c_n\}$. Also in a nonparametric setup, DasGupta, Ghosh and Zen (1995) provide a method for constructing multivariate confidence sets under some shape restrictions (star-shapedness) on the density level sets.

This concept of α-outlier can be rephrased in the language of nonparametric statistical quality control. Suppose a production line is being monitored and we want to decide whether an observation Z corresponding to a new item matches quality specifications. If f (or f_n when f is unknown) represents the distribution of acceptable items, then we may decide that the process is out of control whenever the new observation Z is an α-outlier. Since this application is the main interest of this paper, we will review some results connecting nonparametric set estimation and quality control in the following section.

Nonparametric multidimensional control charts

Consider a manufacturing process where items with a d-dimensional quality characteristic X are produced. Assume that, when the process is in control, X follows a probability density f on $\mathbb{R}^d$. In order to check the stability of the production process some items are sequentially drawn. Assume that the first n observations $X_1,\ldots,X_n$ are independent and are taken while the process is still in

control. What we want to decide is whether or not a new observation X_{n+1} comes also from f. That is, we have to decide if the system has gone "out of control" at stage $n+1$, in the sense that the distribution of X_{n+1} is different from that of the previous observations $X_1, \ldots, X_n$.

We can restate this problem in the language of sequential change-point detection (see, e.g. Yakir 1998) and then $n+1$ would be a possible change-point in the distribution of the process. In either of these frameworks the procedures suggested to solve the problem in the multidimensional setting have in general been of a parametric nature. This is the case of Shewhart, Cusum and Shiryaev-Roberts procedures, for example (see Montgomery 2001 for a review of these methods). These approaches are based on tolerance regions (from which the typical control charts are derived) and constructed in order to control the false alarm probability.

In the unidimensional setting ($d = 1$) the proposed nonparametric schemes are different in nature. Usually the signs or the signed ranks of the observations serve to construct a sequence of statistics and again a change is reported whenever the sequence exceeds a critical level (a detailed account may be found in Csörgo and Horváth 1997 ; see also Gordon and Pollak 1994). An extension of this procedure to multivariate observations demands first a way of ranking them. This has been done (see, e.g. Liu 1995 and references therein) by defining the depth of each datum with respect to the "central core" of the distribution. Thus the problem is reduced to the construction a univariate control chart made from these data depths, but the chart depends on the definition of depth used (simplicial, halfplane, ...)

We are interested in a detection scheme first proposed by Devroye and Wise (1980), which is based on the use of a set estimator. More specifically, the proposal is as follows : if we have an estimator, S_n, of the support of f or of the level set $\{f > c\}$ (for c sufficiently small this can be considered as the "significant support" of f), we decide that there is a change in the distribution at stage $n+1$ if

$$X_{n+1} \notin S_n. \tag{5}$$

The set estimator S_n will depend on some parameter (the smoothing parameter ϵ_n in (1) or the level c in (2)), which can be fixed in advance according to different criteria or chosen in order to control the probability of false alarm

$$P_f\{X_{n+1} \notin S_n\} = \int_{S_n^c} f \tag{6}$$

(like in the classical Shewhart control chart). Each of the following subsections will focus on a different set estimator ((1) or (2)) and mention some of the results attained for that particular choice.

→ Devroye and Wise estimator. This estimator was introduced by Devroye and Wise (1980) in the context of change-point detection schemes. However "naive"

it may seem, it has several advantages over more sophisticated estimators, for example, its computational simplicity or the capability of incorporating shape restrictions via an adequate choice of ϵ_n (see, for instance, Baíllo and Cuevas (2001) where $S_n(\epsilon_n)$ is star-shaped).

The work of Baíllo, Cuevas and Justel (2000) is devoted to the study of the nonparametric control chart associated to estimator (1). Via McDiarmid's (1989) inequality they obtain convergence rates to zero for the probability of false alarm (6). They also consider the problem of support estimation under the assumption that S is connected, quite a natural shape restriction in the quality control framework. This restriction can be easily translated to estimator (1) by suitably choosing the smoothing parameter ϵ_n. The resulting $\hat{S}_n$, a sort of "connected hull" of the sample, is proved to be consistent with respect to appropriate versions of the pseudodistance d_μ. The proof involves results by Tabakis (1996) and Penrose (1999) on the theory of random trees.

In Baíllo, Cuevas and Justel (2000) there are two proposals for the automatic (data-driven) choice of the smoothing parameter ϵ_n, when the false alarm probability is controlled. They are based on resampling ideas (cross-validation and smoothed bootstrap).

$\rightarrow$ The plug-in estimator. We may consider the performance of estimator (2) in two contexts. In the context of quality control we will decide that the process is out of control if observation X_{n+1} belongs to $\{f_n < c\}$. In the clustering setting, the clusters in the population given by density f could be estimated by each of the connected components of $\{f_n > c\}$. This means that observatins falling outside this set are disregarded in the sense that they are not classified. In both situations it is interesting to know how fast the "probability content" of the empirical level set $\{f_n < c\}$ approaches its population limit.

In particular, Baíllo, Cuesta-Albertos and Cuevas (2001) obtain L^1-convergence rates for $P_n(Z)$, where $P_n(z)$ is the probability of not classifying datum z and Z is an observation which follows density f. On the other hand, they consider the case where level c is chosen as in (4). This choice is reasonable in the context of quality control when the desired false alarm probability is α. These authors obtain rates of convergence to α of the real false alarm probability (6).

Références

- Baíllo, A., Cuesta-Albertos, J. A. and Cuevas, A. (2001). Convergence rates in nonparametric estimation of level sets. *Stat. Prob. Letters* 53, 27-35.
- Baíllo, A. and Cuevas, A. (2001), On the estimation of a star-shaped set.

Adv. Appl. Prob. (SGSA), 33, 717-726.
– Baíllo, A., Cuevas, A. and Justel, A. (2000). Set estimation and nonparametric detection. *Canad. J. Statist.*, 28, 765-782.
– Bertholet, V., Rasson, J. P. and Lissoir, S. (1998). About the automatic detection of training sets for multispectral images classification. *Advances in Data Science and Classification*, pp. 221–226. Springer.
– Chevalier, J. (1976). Estimation du support et du contour de support d'une loi de probabilité. *Ann. Inst. H. Poincaré, sec. B*, 12, 339–364.
– Csörgo, M. and Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley.
– Cuevas, A. (1990). On pattern analysis in the non-convex case. *Kybernetes*, 19, 26–33.
– Cuevas, A. and Fraiman, R. (1997). A plug-in approach to support estimation. *Ann. Statist.*, 25, 2300–2312.
– Cuevas, A., Febrero, M. and Fraiman, R. (2000). Estimating the number of clusters. *Canad. J. Statist.*, 28, 2, 367-382.
– DasGupta, A., Ghosh, J. K. and Zen, M. M. (1995). A new general method for constructing confidence sets in arbitrary dimensions : with applications. *Ann. Statist.*, 23, 1408-1432.
– Devroye, L. and Wise, G. L. (1980). Detection of abnormal behavior via nonparametric estimation of the support. *SIAM J. Appl. Math.*, 38, 480-488.
– Dümbgen, L. and Walther, G. (1996). Rates of convergence for random approximations of convex sets. *Adv. Appl. Prob*, 28, 384–393.
– Efron, B. (1965). The convex hull of a random set of points. *Biometrika*, 52, 331–343.
– Geffroy, J. (1964). Sur un problème d'estimation géométrique. *Publications de l'Institut de Statistique des Universités de Paris*, 13, 191-210.
– Gordon, L. and Pollak, M. (1994). An efficient sequential nonparametric scheme for detecting a change of distribution. *Ann. Stat.*, 22, 2, 763-804.
– Grenander, U. (1981). *Abstract Inference*. Wiley.
– Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley.
– Korostelev, A. P. and Tsybakov, A. B. (1993). *Minimax Theory of Image Reconstruction*. Springer-Verlag.
– Liu, R. (1995). Control charts for multivariate processes. *J. Amer. Stat. Assoc.* 90, 1380–1387.
– McDiarmid, C. (1989). On the method of bounded differences. LMS Lecture Notes Series, 141, 148-188. Cambridge University Press, Cambridge.
– Molchanov, I. S. (1998). A limit theorem for solutions of inequalities. *Scand. J. Statist.*, 25, 235-242.
– Montgomery, D. C. (2001). *Introduction to Statistical Quality Control*. Wiley.
– Moore, M. (1984). On the estimation of a convex set. *Ann. Statist.*, 12,

1090-1099.
– Penrose, M. D. (1999). A strong law for the longest edge of the minimal spanning tree. *Ann. Probab.*, 27, 1, 246-260.
– Pollak, M. and Siegmund, D. (1991). Sequential detecting of a change in a normal mean when the initial value is unknown. *Ann. Statist.*, 19, 394-416.
– Polonik, W. (1995). Measuring mass concentrations and estimating density contour clusters. An excess mass approach. *Ann. Statist.*, 23, 855-882.
– Ramsay, J. O. and Silverman, B. W. (1997). *Functional Data Analysis*. Springer.
– Rényi, A. and Sulanke, R. (1963). Über die konvexe Hülle von n zufällig gewählten Punkten. *Z. Wahrscheinlichkeitsth. verw. Geb.*, 2, 75-84.
– Rudemo, M. and Stryhn, H. (1994). Boundary estimation for star-shaped objects. *Change-point problems*. IMS, California, 276-283.
– Schneider, R. (1988). Random approximation of convex sets. *J. Microscopy*, 151, 211–227.
– Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer-Verlag.
– Stoyan, D., Kendall, W. S. and Mecke, J. (1995). *Stochastic Geometry and its Applications*. 2nd. edition. Wiley.
– Tabakis, E. (1996). On the longest edge of the minimal spanning tree. *From Data to Knowledge*, pp. 222–230. Springer-Verlag.
– Tsybakov, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.*, 25, 948-969.
– Walther, G. (1997). Granulometric smoothing. *Ann. Statist.*, 25, 2273-2299.
– Yakir, B. (1998). On the average run length to false alarm in surveillance problems which possess an invariance structure. *Ann. Statist.*, 26, 1198-1214.

.

ACP d'un processus stationnaire sous contrainte de B-mesurabilité. Lien avec le produit tensoriel de mesures spectrales

Tawfik BENCHIKH* et Abderrahmane YOUSFATE
* Adresse pour correspondance :
Laboratoire de Mathématiques.
UDL, Sidi bel Abbès 22000, Algérie.
e-mail : benchikh@univ-sba.dz

Résumé

Dans ce travail nous présentons l'*A.C.P.* "pas-à-pas" d'un opérateur linéaire continue U sous contrainte de mesurabilité par rapport à une tribu $\mathcal{B}$. Nous présentons d'abord les résultats généraux de l'*A.C.P.* sous contrainte linéaire, ensuite nous présentons les résultats quand l'opérateur U est associé à un processus stationnaire de second ordre. Dans le cas fini, l'*A.C.P.* sous contrainte linéaire a connu plusieurs axes selon les critères d'optimisations : choix de métrique ([11]), variables instrumentales ([9]), ... Dans le cas infini on trouve les travaux de [10] (ACPCL sur variables hilbertiennes), [12] (ACP et filtrage), [2] (Solutions splines), [8] (*A.C.P.C.L.* sur les L^p), [1] (*A.C.P.C.L.* sur des espaces de Banach). Enfin, tenant compte de la décomposition spectrale de U dans le cadre de l'*A.C.P.C.L.* et du calcul opératoriel sur les mesures spectrales, un essai de comparaison des approches est abordé.

1. Introduction.

Soient H et H' deux espaces de Hilbert séparables et soit U un opérateur de $\mathcal{L}(H, H')$. Faire l'*A.C.P.* de U revient à faire la décomposition spectrale de l'opérateur $U^\star \circ U$ noté W ou celle de $U \circ U^\star$ noté V. Pour tout complément de détails sur l'*A.C.P.* voir [6]. Pour faire l'*A.C.P.C.L.* (généralisant les travaux de ([11]) et de ([9]) on utilise la définition suivante.

Définition
Soit G un sous espace vectoriel fermé de H. On appelle $A.C.P.C.L_1$ "pas-à-pas" de

l'opérateur U sous la contrainte G, la méthode itérative d'optimisation suivante :

$$\begin{cases} \max\limits_{w \in G_1} \|Uw\| \\ \\ \|w\|_H = 1; \end{cases} \tag{7}$$

si w_1 est argument de la solution, le deuxième argument w_2 doit vérifier

$$\begin{cases} \max\limits_{w \in G_1} \|Uw\| \\ \\ \|w\|_H = 1 \\ \\ w \perp w_1; \end{cases} \tag{8}$$

et itération sous contrainte d'orthonormalité des arguments des solutions.

En prenant G un sous espace vectoriel fermé de H et P le projecteur orthogonal dans H d'image G, on remarque que L'$A.C.P.C.L_1$ "pas-à-pas" de l'opérateur U sous G est l'$A.C.P.$ "pas-à-pas" de l'opérateur $U \circ P$ de $\mathcal{L}(H, H')$.

Proposition
L'$A.C.P.C.L.1$ de l'opérateur U sous G s'obtient par l'analyse spectrale de l'opérateur $P \circ W \circ P$.

De même que pour l'$A.C.P.C.L._1$, on peut déduire que si G' est un sous espace vectoriel fermé de H' et si P' est le projecteur orthogonal dans H' d'image G', L'$A.C.P.C.L.$ "pas-à-pas" de l'opérateur U sous G' est l'$A.C.P.$ "pas-à-pas" de l'opérateur $P' \circ U$ de $\mathcal{L}(\mathcal{H}, \mathcal{H}')$. Cette $A.C.P.C.L.$ est appelée $A.C.P.C.L._2$. On peut facilement déduire de l'$A.C.P.C.L._1$ que l'$A.C.P.C.L._2$ sous G' de l'opérateur U s'obtient par l'analyse spectrale de l'opérateur $P' \circ V \circ P'$.

En général l'$A.C.P.C.L._1$ et l'$A.C.P.C.L._2$ sont deux problèmes qui donnent des résultats non équivalents, contrairement à l'$A.C.P.$ qui peut être obtenue aussi bien à partir de $U^\star \circ U$ que de $U \circ U^\star$.

Nous pouvons également introduire simultanément l'$A.C.P.C.L._1$ et l'$A.C.P.C.L._2$. Ainsi l'$A.C.P.C.L.$ simultanée revient à l'$A.C.P.$ de l'opérateur $P' \circ U \circ P$.

Dans le cas général, la méthode "pas-à-pas" s'arrête au premier point d'accumulation du spectre de l'opérateur de covariance en suivant l'ordre décroissant des valeurs propres. Aussi allons-nous définir, par la suite, l'$A.C.P.$ ou l'$A.C.P.C.L.$ sur l'ensemble du spectre de l'opérateur de covariance issu de l'opérateur considéré. Cette présentation sera qualifiée de globale.

Au cas où H et H' sont de dimensions finies, on montre facilement que l'$A.C.P.$ "pas-à-pas" est une $A.C.P.$ totale et épuise, par conséquent, toutes les solutions de l'$A.C.P.$ globale. Dans cette situation l'expression des projecteurs est aisée et numériquement calculable.

Par la suite, on considère $\mathcal{P} = \{P, P \in \mathcal{L}(H), P \circ P = P\}$ et $\mathcal{P}' = \{P', P' \in \mathcal{L}(H'); P' \circ P' = P'\}$.

Proposition
Pour tout $(P, P') \in \mathcal{P} \times \mathcal{P}'$, l'endomorphisme

$$\begin{array}{rccc} \mathbf{\Pi_{P,P'}}: & \mathcal{L}(H,H') & \to & \mathcal{L}(H,H') \\ & U & \longmapsto & P' \circ U \circ P \end{array}$$

est un projecteur dans $\mathcal{L}(H, H')$.

En notant $\mathcal{A}(H) = \{A \in \mathcal{L}(H); A = A^*\}$, on constate que $\Pi_{P,P}\mathcal{A}(H) \subset \mathcal{A}(H)$. De même si $\mathcal{A}_+(H) = \{A \in \mathcal{A}(H); A \text{ non négatif}\}$, on a $\Pi_{P,P}\mathcal{A}_+(H) \subset \mathcal{A}_+(H)$.

Remarques

1. Si $P' = I$, l'ACP de $\Pi_{P,P'}(U)$ c'est l'$ACPCL_1$. Ainsi l'opérateur à décomposer s'écrit $P \circ U^* \circ U \circ P = \Pi_{P,P}(W)$ où $\Pi_{P,P} \in \mathcal{L}(\mathcal{L}(H))$.
2. Si $P = I$, l'ACP de $\Pi_{P,P'}(U)$ c'est l'$ACPCL_2$, alors $P' \circ V \circ P' = \Pi_{P',P'}(V)$ où $\Pi_{P',P'} \in \mathcal{L}(\mathcal{L}(H'))$.
3. Dans le cas général, l'ACP de $\Pi_{P,P'}(U)$ c'est l'$ACPCL$ mixte et l'opérateur à décomposer s'écrit $\Pi_{P,P'}(U \circ P \circ U^*) = P' \circ U \circ P \circ U^* \circ P'$. Si P commute avec U $(H = H')$ alors on a à décomposer $\Pi_{P',P'}\Pi_{P,P}(U \circ U^*)$. Si de plus $U \circ U^* \in \Pi_{P',P'}(\mathcal{L}(\mathcal{L}(H')) = \mathcal{L}(\mathcal{L}(H))$, la décomposition spectrale nous rappelle une certaine approche de l'analyse canonique.

2. A.C.P.C.L. de processus aléatoire.

Rappelons qu'un processus $(X_t)_{t\in T}$ de second ordre (la variance de chaque $X_t,\ t \in T$ est finie) admet une $A.C.P.$ associée à l'opérateur

$$U : \mathcal{L}^2(\Omega) \longrightarrow \mathcal{L}^2(T)$$

$$Uh = \int_\Omega X_t(\omega)h(\omega)d\mathbb{P}(\omega)$$

d'où

$$U^* : \mathcal{L}^2(T) \longrightarrow \mathcal{L}^2(\Omega)$$

$$U^* = \int_T g(t)X_t = {}^t X g.$$

L'espace vectoriel $\mathcal{L}^2(\Omega)$ est généré par les X_t ; $t \in T$ et l'espace vectoriel $\mathcal{L}^2(T)$ est généré par les X_ω ; $\omega \in \Omega$. On suppose aussi que $E(X_t) = 0$.

Pour faire l'ACP de $(X_t)_{t\in T}$, il suffit de faire la décomposition spectrale de l'opérateur $U \circ U^*$ (qui est un opérateur de covariance) ou celle de $U^* \circ U$ (qu'on appelle opérateur d'Escoufier). L'$A.C.P$ de U permet la décomposition spectrale du processus (X_t) sous forme d'une somme de processus quasi-déterministes du type :

$$X_t = \sum_{k=1}^{\infty} \sqrt{\lambda_i}\xi_k f_k(t) \qquad m.q.$$

où les ξ_k sont des variables aléatoires normées ne dépendant pas du temps et les $f_k(t)$ des fonctions déterministes normées exprimées en fonction du temps. Cette décomposition est possible même en cas de non stationarité du processus. Dans le cas stationnaire, nous avons le théorème important qui suit.

Proposition (Décomposition de Karhunen-Loève).
Pour que $X_t = \sum_{k=1}^{\infty} \sqrt{\lambda_i}\xi_k f_k(t)$ $m.q.$ où les $f_k(t)$ et les ξ_k forment des systèmes orthonormaux respectifs de $L^2(\mathrm{T})$ et de $L^2(\Omega)$, il faut que les f_k soient un système de fonctions propres de l'opérateur de covariance V du processus et les λ_i les valeurs propres correspondantes.

Nous avons également si $(X_t)_{t\in T}$, $T = \mathbb{R}$, est un processus stationnaire au sens large, alors $(X_t)_{t\in T}$ admet la représentation de Cramer :

$$X_t = \int_{-\infty}^{+\infty} e^{it\lambda} dZ(\lambda)$$

où $Z(\lambda)$ est un processus complexe à accroissements orthogonaux. La question qui survient de manière naturelle est : "Y a-t-il un lien entre les deux décompositions spectrales ?" Faire l'$A.C.P.C.L.$ sur un tel processus revient à faire une $A.C.P.$ tout en introduisant une projection sur $\mathcal{L}^2(\Omega)$, sur $\mathcal{L}^2(T)$ ou sur les deux simultanément.
Soit maintenant $\mathcal{B}$ une sous tribu complète de $\mathcal{A}$. Notons alors $G = \{Y_t \in \mathcal{L}^2(\Omega); Y_i = E^{\mathcal{B}}(X_t)\}$ où $E^{\mathcal{B}}$ est l'espérance conditionnelle relativement à $\mathcal{B}$ et $U_{\mathcal{B}}$ la restriction de U àG que l'on note U/G. Comme $E^{\mathcal{B}}$ est un projecteur dans $\mathcal{L}^2(\Omega)$, on a $U_{\mathcal{B}} = U \circ E^{\mathcal{B}}$.

Proposition
L'ACP de U sous contrainte de $\mathcal{B}$-mesurabilité des composantes principales re-

vient à faire l'ACP de $U_{\mathcal{B}}$.

Remarque.

- L'ACP de $U_{\mathcal{B}}$ revient à faire la décomposition spectrale de l'opérateur $U_{\mathcal{B}} \circ U_{\mathcal{B}}^* = UE^{\mathcal{B}}U^*$ ou celle de $U_{\mathcal{B}}^* \circ U_{\mathcal{B}} = E^{\mathcal{B}} \circ U^* \circ U \circ E^{\mathcal{B}}$ et l'on retrouve la démarche de l'$ACPCL_1$ où P' est remplacé par $E^{\mathcal{B}}$.
- L'ACP "pas à pas" de U est totale ssi le seul point d'accumulation de $Sp(U \circ U^*)$ est la borne inférieure de $Sp(U \circ U^*)$.
- Si l'ACP "pas à pas" de U n'est pas totale et si la plus grande valeur propre est de multiplicité finie, pour tout λ point d'accumulation de $Sp(U \circ U^*)$, il existe une tribu complète $\mathcal{B}$ tel que $E^{\mathcal{B}} \circ (U \circ U^* - \lambda I) \circ E^{\mathcal{B}}$ soit compact. pas totale, chaque point d'accumulation de
- Si l'ACP " pas à pas " de U est totale telle que $U \circ U^*$ soit inversible, alors $I = (U \circ U^*)^{-1} \circ U \circ E^{\mathcal{B}} \circ U^* + (U \circ U^*)^{-1} \circ U \circ (I - E^{\mathcal{B}}) \circ U^*$. Ainsi l'$ACP$ de U sous la contrainte que les composantes principales soient $\mathcal{B}-$mesurables tel que E soit muni de la métrique $(U \circ U^*)^{-1}$ revient à faire l'analyse discriminante de U sous $\mathcal{B}$. La technique est d'autant plus intéressante en classification quand $\mathcal{B}$ est engendrée par une partition.

Proposition
Soient $\mathcal{B}_1$ et $\mathcal{B}_2$ deux sous-tribus complète de $\mathcal{A}$, alors
$E^{\mathcal{B}_1 \cap \mathcal{B}_2} = E^{\mathcal{B}_1} \circ E^{\mathcal{B}_2} = E^{\mathcal{B}_2} \circ E^{\mathcal{B}_1}$. Donc l'$ACP$ conditionnelle de U par rapport à la tribu $\mathcal{B}$ engendrée par $\mathcal{B}_1$ et B_2 est l'ACP de $U \circ E^{\mathcal{B}}$ où

$$E^{\mathcal{B}} = E^{\mathcal{B}_1} + E^{\mathcal{B}_2} - E^{\mathcal{B}_1} \circ E^{\mathcal{B}_2} = E^{\mathcal{B}_1}(I - E^{\mathcal{B}_2}) + E^{\mathcal{B}_2}(I - E^{\mathcal{B}_1}) + E^{\mathcal{B}_1}E^{\mathcal{B}_2}.$$

En notant $E^{\mathcal{B}_1} \circ (I - E^{\mathcal{B}_2}) = E^{\mathcal{B}_1^-}$, $E^{\mathcal{B}_2} \circ (I - E^{\mathcal{B}_1}) = E^{\mathcal{B}_2^-}$, $E^{\mathcal{B}_1} \circ E^{\mathcal{B}_2} = E^{\mathcal{B}_{12}}$ et $E^{\mathcal{B}_{12}^-} = E^{\mathcal{B}_1^-} + E^{\mathcal{B}_1^-}$, on remarque que $E^{\mathcal{B}_1^-}$, $E^{\mathcal{B}_2^-}$ et $E^{\mathcal{B}_{12}}$ sont des projecteurs dont les produits deux à deux sont nuls. Ainsi, on constate que nous sommes en train de préparer le terrain pour une décomposition spectrale liée à la tribu engendrée par l'union des tribus $\mathcal{B}_1$ et $\mathcal{B}_2$.

En notant $\mathcal{C}$ une sous-tribu complète de $\mathcal{T}$ et $E^{\mathcal{C}}$ l'espérance conditionnelle associée, nous construisons G' un $s.e.v.$ fermé de $\mathcal{L}^2(T)$ tel que $E^{\mathcal{C}}(\mathcal{L}^2(T)) = G'$. Cest une façon de considérer que l'$A.C.P.$ de U est sous la contrainte que les axes principaux doivent être contenus dans G'. Ce qui revient à faire l'ACP de U' tel que :

$$\begin{array}{ccc} U'^* : \mathcal{L}^2(\Omega) & \longrightarrow & \mathcal{L}^2(T) \\ g & \longmapsto & U^* \circ E^{\mathcal{C}} g \end{array}$$

On remarque que $U'^* = U^*/G'$. On peut donc substituer U' par $E^{\mathcal{C}} \circ U$. Cela permet d'énoncer la proposition qui suit.

Proposition
L'ACP de U sous contrainte de $\mathcal{C}$-mesurabilité des axes principaux est l'ACP de $E^{\mathcal{C}} \circ U$.

Nous pouvons mettre également une double condition sur $\mathcal{L}^2(\Omega)$ et $\mathcal{L}^2(T)$ simultanément. A ce moment L'ACP conditionnelle sous $G' \subset \mathcal{L}^2(T)$ et sous $\mathcal{B}$-mesurabilité des composantes principales revient à faire l'ACP de $E^{\mathcal{C}} \circ U \circ E^{\mathcal{B}}$.

3. Mesure aléatoire spectrale associée à un processus stationnaire

Soit $(X_t)_{t\in T}$ un processus stationnaire de second ordre ; il existe alors un unique processus Z à accroissements orthogonaux vérifiant

$$X_t = \int_{\mathbb{R}} e^{it\lambda} dZ(\lambda)$$

où Z est une application définie sur $\mathcal{B}$ à valeurs dans $L^2(\Omega)$ que l'on va noter H par la suite. Elle est appelée mesure aléatoire $(m.a)$. Elle vérifie pour tout couple (A, B) d'éléments disjoints de $\mathcal{B} \times \mathcal{B}$:

$$Z(A \cup B) = Z(A) + Z(B) \text{ et } < Z(A), Z(B) >= 0.$$

En considérant que $T = [-a, a]$, il suffit de décomposer $e^{it\lambda}$ en série de sinus pour exprimer le théorème de Shannon

$$X_t = \sum_{n\in\mathbb{Z}} \frac{sin(at - \pi n)}{at - \pi n} X_{\pi n/a} \quad p.s.$$

De la même manière, dans le domaine des fréquences, en considérant que $\Omega = [-b, b]$, on peut écrire

$$X_\omega = \sum_{n\in\mathbb{Z}} \frac{sin(b\omega - \pi n)}{b\omega - \pi n} X_{\pi n/b} \quad p.s.$$

En utilisant les techniques de l'$A.C.P.$, on déduit que la mesure aléatoire induit une décomposition de $L^2(\Omega)$ à laquelle on peut associer une famille de projecteurs qui commutent entre eux. Ainsi on peut construire

$$\begin{array}{rccl} \varepsilon : & \mathcal{B} & \to & \mathcal{P} \\ & A & \longmapsto & \varepsilon(A) : \begin{array}{ccc} H & \to & H \\ X & \longmapsto & Z^X(A) \end{array} \end{array}$$

où Z^X est une mesure aléatoire vérifiant $Z^X(\Omega) = X$ et pour tout $Y \in L^2(\Omega)$, on a Z^X et Z^Y stationnairement corrélés (voir [3] et [4] pour plus de détails.)

Proposition
Pour tout couple (A, B) élément de $\mathcal{B} \times \mathcal{B}$: $\varepsilon(A) \circ \varepsilon(B) = \varepsilon(A \cap B)$.

Par la suite, on définit le produit tensoriel de deux mesures spectrales ε_1 et ε_2 comme suit

$$\begin{array}{rccc} \varepsilon_1 \otimes \varepsilon_2 : & \mathcal{B} \times \mathcal{B} & \longrightarrow & \mathcal{P} \\ & (A, B) & \longmapsto & \varepsilon_1(A) \circ \varepsilon_2(B) = \varepsilon_2(B) \circ \varepsilon_1(A) \end{array}$$

Cette définition n'a de sens que si $\varepsilon_1(A)$ et $\varepsilon_2(B)$ commutent.

Si ε_1 agit sur H et ε_2 agit sur H', pour tout $(A, B) \in \mathcal{B}^2$, le produit tensoriel fonctionnel des deux mesures spectrales s'écrit :

$$\begin{array}{rccc} \varepsilon_1(A) \overset{l}{\otimes} \varepsilon_2(B) : & \mathcal{L}(H, H') & \longrightarrow & \mathcal{L}(H, H') \\ & U & \longmapsto & \varepsilon_2(B) \circ U \circ \varepsilon_1(A) \end{array}$$

et l'on retrouve les expressions à analyser dans le cadre de l'*A.C.P.C.L.* Si U est un endomorphisme, une analyse spécifique est faite pour les opérateurs unitaires.

References

1 T. Benchikh, (1999). Analyses factorielles dans un espace de Banach sous contraintes linéaires", Magister, Sidi-Bel-Abbès.

2 P. Besse, (1989). Approximation spline et optimalité en analyse en composantes principales, Thèse de doctorat d'état es-sciences mathématiques, Toulouse III.

3 A. Boudou, (2000). Mesure spectrales et Processus Stationnaire.

4 A. Boudou et Y. Romain, (2000). On spectral and random measures associated to discrete and continus-time processes.

5 H. Brezis, (1987). *Analyse Fonctionnelle, Théorie et Applications*, 2^{e} édition, Edition MASSON, Paris.

6 J. Dauxois et A. Pousse, (1976). Les analyses factorielles en calcul des probabilité et en statistique : essai d'etude synthétique , Thèse es-sciences , Toulouse.

7 J. Durand, (1992). "Generalized principal component analysis with respect to instrumental variables via univariate spline transformations, *Computational statistics and data analysis*, **16**, p. 423-440.

8 F. Rachedi, (1996). "ACPCL dans un espace de Hilbert et essai d'extension dans un espace de Banach, Magister, Sidi-Bel-Abbès.

9 R. Sabatier, (1987). Méthodes factorielles en analyse des données : Approximations et prise en compte des variables concomitantes, Thèse Es-Sciences, Montpellier.

10 S. Smahdi et A. Yousfate, (1992). Extension de l'ACPCL à une variable hilbertienne", 3^{rd} islamic countries conference on statistical sciences approaching, Rabat.

11 F.O. Tebboune, (1994). L'analyse en composantes principales sous contraintes linéaires, *Annales de mathématique, Université de Sidi-Bel-Abbès*, **1**, pp 29-38.

12 A. Yousfate, (1992). A.C.P. et filtrage, MOAD'92, Béjaïa.

Recalage de courbes et analyse de variance fonctionnelle par ondelettes

Jérémie BIGOT
Laboratoire IMAG-LMC, Université Joseph Fourier
BP 53, 38041 Grenoble Cedex 9, France
e-mail : Jeremie.Bigot@imag.fr

Résumé

Lors de l'étude d'un processus chez différents individus, les courbes observées présentent généralement des caractéristiques similaires. Un problème important consiste alors à déterminer l'allure caractéristique de ce processus ou bien à tester s'il existe des différences significatives parmi deux sous-ensembles de sujets. L'estimation des variations au sein d'une population à partir de données fonctionnelles, peut se résoudre à l'aide de l'*Analyse de variance fonctionnelle* (FANOVA) qui généralise au cas de fonctions, les techniques d'*Analyse de variance* (ANOVA) [1],[4]. Toutefois, afin de comparer des objets qui présentent des caractéristiques similaires, il est nécessaire de trouver un système référentiel commun pour les représenter [2]. En effet, la présence de bruit et l'existence de variations temporelles entre les courbes font que la simple moyenne des fonctions observées n'est pas, en général, un bon estimateur de la forme typique d'une courbe. Une approche possible consiste alors à trouver, pour chaque courbe observée, une fonction de déformation afin de synchroniser l'ensemble des courbes avant d'en faire la moyenne ou d'appliquer n'importe quelle autre procédure d'inférence statistique. L'alignement de deux fonctions peut se faire à partir de la position de leurs points caractéristiques (ou landmarks e.g. extrema, points d'inflexion, singularités) [5], [4]. Une telle méthode dépend donc fortement de la qualité de l'estimation de ces derniers. Les ondelettes constituent un outil très puissant pour la caractérisation de la structure locale d'un signal [3]. En particulier, il est possible de détecter les points caractéristiques d'une fonction en suivant la propagation des zero-crossings de sa transformée en ondelettes quand le niveau de résolution augmente. Une approche non paramétrique est proposée pour estimer les zero-crossings de la transformée en ondelettes d'un signal bruité à différentes échelles. Afin d'éliminer les erreurs d'estimation dues au bruit, un nouvel outil est introduit qui calcule la "densité" ou "intensité structurelle" des zero-crossings. Les modes de cette "densité" correspondent alors à la position des landmarks

du signal. Une des principales difficultés des méthodes de recalage à partir de landmarks est la détermination des paires de points caractéristiques qui doivent être mis en correspondance. Il est montré que l'intensité structurelle permet de définir une nouvelle méthode de recalage automatique de courbes à partir de leurs landmarks. Quelques simulations et une application à des données bio-médicales servent d'illustrations des méthodes proposées.

Références

1 Abramovich, F. Antoniadis, A. Sapatinas, T. et Vidakovic, B. (2002). Optimal Testing in Functional Analysis of Variance, *To appear in JASA*.

2 Kneip, A. et Gasser, Th. (1992). Statistical tools to analyze data representing a sample of curves, *Ann. Statist.*, **20** No.3 1266-1305.

3 Mallat, S. et Hwang, W.L. (1992). Singularity Detection and Processing with Wavelets, *IEEE Trans. Inform. Theory*, **38** No.2 617-643.

4 Ramsay, J.O. et Silverman, B.W. (1997). Functional data analysis, *New York : Springer Verlag.*

5 Younes, L. (2000). Deformations, Warping and Object Comparison, *Tutorial*,
`http ://www.cmla.ens-cachan.fr/ỹounes`.

Modèle mixte de régression multiple à coefficients lisses : application à l'étude de captures journalières de civelles d'Anguille dans l'estuaire de l'Adour en fonction de variables environnementales

Anestis ANTONIADIS et Noëlle BRU *
* Adresse pour correspondance :
Université Pierre Mendès-France,
LabSAD, Grenoble.
e-mail : Noelle.Bru@upmf-grenoble.fr

Résumé

Nous nous proposons de nous intéresser à une classe particulière de modèles, le modèle de régression multiple à coefficients lisses avec effets mixtes. Notre intérêt s'est porté sur l'étude de ce type de modèle afin de répondre à l'analyse d'un jeu de données relatif à l'étude de captures journalières de civelles d'Anguille dans l'estuaire de l'Adour sur différentes saisons de pêche en fonction de variables environnementales.

Ce travail effectué en collaboration avec le laboratoire halieutique IFREMER d'Aquitaine vise, entre autre, à répondre à deux questions principales :

- extraire une information pertinente sur la dynamique de ce phénomène en fonction de covariables plus ou moins influentes à exhiber ;
- prévoir l'évolution de la saison de pêche suivante en proposant différents scénarios basés sur la simulation de conditions environnementales.

Ce travail passe par un aspect de construction de modèles qui soient à la fois parcimonieux et simple d'interprétation et avec un effet prédictif certain.

Le modèle de régression linéaire multiple à coefficients lisses avec effets mixtes a été choisi parce qu'il correspond aux critères mis en avant ci-dessus, à savoir :

1. la fonction de régression est linéaire donc facile à interpréter ;
2. l'introduction d'effets mixtes permet de prendre en compte un effet fixe relatif à une évolution moyenne du phénomène sur l'ensemble des saisons qui peut être interpréter comme une évolution commune du phénomène en terme d'abondance des poissons et un effet alétaoire propre à chaque saison qui dépend des conditions environnementales spécifiques à chaque période ;

3. Le lissage intervient en fin de méthodologie permettant ainsi de ne pas biaiser l'information contenue dans les données brutes et de ne pas fausser la vision du phénomène. Plusieurs types de lissage peuvent ainsi être comparés une fois l'information extraite des données brutes. Le lissage des coefficients permet : d'une part de comparer l'influence de chaque variable jour par jour sur l'ensemble des saisons de telle sorte à pouvoir considérer des observations temporelles non identiques et non équidistantes et d'autre part de faire également apparaître une certaine dépendance temporelle à l'intérieur des saisons de pêche.

Comparison of parametric and semiparametric estimates in a degradation model with covariates and traumatic censoring

Vincent COUALLIER
Équipe Statistique Mathématique et ses Applications
U.F.R. Sciences et Modélisation,
Université Victor Segalen Bordeaux II
146 rue Leo Saignat 33076 Bordeaux cedex
e-mail : couallier@sm.u-bordeaux2.fr

Abstract

There exists numerous ways to construct a statistical model for the evolution of the degradation of a subject. The most important models include finite state Markov chains or semi-Markov processes (for which some transitions means deteriorating or improving and where the states reflect the levels of degradation) or models describing the degradation by the evolution (often increasing) of a random function of time. In this last case, the individual trajectories can themselves be modelled by some known parametric function with random coefficients or by the paths of a general stochastic process. This process has to reflect the way degradation evolves ; in some cases, it increases slowly and gradually, in other cases it results from shocks, each of them leading to an increment of degradation.

The main aim of this work is to analyse and compare the estimation in a specific statistical degradation model studied in Bagdonavicius & Nikulin, (2001), with parametric or semi-parametric assumptions. Computational aspects of the maximum likelihood or pseudo maximum likelihood estimates are discussed. A second aim of the paper is to address the computational issues which arise in such complex degradation model with covariates and censored data.

The degradation model

As in Bagdonavicius & Nikulin, (2001), Singpurwalla, (1995) and others we choose here a model in which the space of degradation levels is not finite nor coun-

table but an interval of $\mathbb{R}$. The degradation process is modelled by an increasing stochastic process (hence wiener diffusion process can not be considered). This process is supposed to verify :

– $EZ(t) = m(t)$ where m is an unknown function.

– Z is a increasing process with independent increments.

We will consider the nonparametric and parametric cases to model the mean function m. In this last case we have

$$EZ(t) = m(t, \theta), \quad t > 0,$$

where m is a known function and $\theta \in I\!R^p$ is a parameter to be estimated.

Following Bagdonavicius & Nikulin, (2000), Z is supposed to lie in the family of Gamma processes with right-continuous trajectories. It ensures the growth of the paths and the fact that, denoting

$$Z(t) = \sigma^2 \gamma(t) \quad \text{with } \gamma(t) \sim G(1, \nu(t)) = G(1, \frac{m(t)}{\sigma^2}),$$

the increment from t to $t + \Delta t$ for $\Delta t > 0$ lies in the same family of distribution,

$$\Delta\gamma(t) = \gamma(t + \Delta t) - \gamma(t) \sim G(1, \frac{\Delta m(t)}{\sigma^2}).$$

A failure caused by degradation occurs when $Z(t)$ reaches a fixed threshold z_0. The moment of failure is

$$T = \sup\{\ t \mid \ Z(t) < z_0 \ \} = \inf\{\ t \mid \ Z(t) \geq z_0 \ \}.$$

Some traumatic events censure the degradation paths

In addition to the breakdown due to wear, it is natural to consider breakdowns whose origin is not directly related to ageing. The probability of these traumatic events can however depends on degradation. Thus, we model the censoring times C due to a traumatic event as the time of first jump of a nonhomogeneous Poisson process $(N(t), t \geq 0)$ whose intensity function can depend on the level of degradation reached at the time t.

Giving the degradation path until time t, we assume that the intensity function depends on $Z(t)$,

$$\lambda(t) = \lambda(Z(t)),$$

and the conditional distribution of the random variable $N(t)$ which is the number of traumatic events until time t is the Poisson distribution with mean $\int_0^t \lambda(Z(s))ds$.

Thus, the conditional survival function of C is

$$Q(t) = \mathbf{P}\{C > t\} = \mathbf{E}\Big\{\exp(-\int_0^t \lambda(Z(s))ds)\Big\}.$$

Whatever the cause of failure is, let us denote $U = \min(T, C)$ the time of failure and G its survival function. We have

$$G(t) = \mathbf{P}\{U > t\} = \mathbf{E}\Big\{\exp(-\int_0^t \lambda(Z(s))ds)\mathbf{1}_{\{Z(t)<z_0\}}\Big\}.$$

In the following, we will assume that the conditional intensity function is a parametric function of $Z(t)$.

Covariates

We model the influence of the environment on the degradation process by covariates in the following way : we have for each item the value $\mathbf{x} = (1, x_1, ..., x_s)$ of the explanatory variable $\mathbf{X} = (1, X_1, ..., X_s) \in I\!R^{s+1}$ assumed to be constant in time. (see Bagdonavicius & Nikulin, 2000 and Singpurwalla , 1995 for dynamic time-varying explanatory variable)

The classical framework of accelerated failure time model can be applied to degradation model by considering that acceleration of time is due to a scaling by means of a function $f_x(t)$ depending on the value of $\mathbf{X}$. We will consider in simulation and implementation of the model that f is linear and X is a categorical multivariate random variable. In this case, we will have, conditionally on $\mathbf{X}$=$\mathbf{x}$,

$$Z_{\mathbf{x}}(t) = \sigma^2\gamma(e^{(\beta^T\mathbf{x})}t) \quad t \geq 0.$$

The parameter $\beta = (\beta_0, \beta_1, .., \beta_s)$ measures the influence of the different covariates. Given $\mathbf{X}$=$\mathbf{x}$, the coefficient $e^{(\beta^T\mathbf{x})}$ is a relative measurement of the (constant) stress of the environment.

The issue is thus to estimate the following reliability charateristics : the mean degradation function m, the mean time needed to attain the threshold z_0, the survival function Q S and G of failures due to degradation only, traumatic event only, or an unspecified cause respectively.

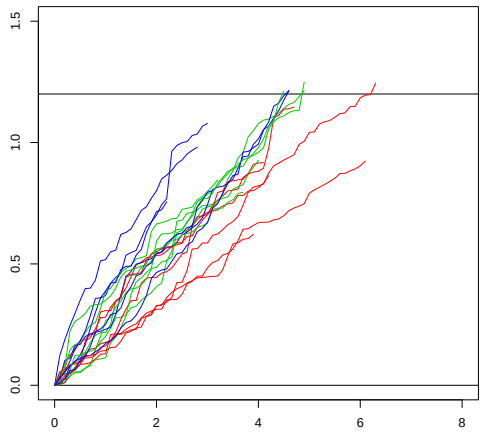

FIG. 1 – 18 paths under 3 different stresses

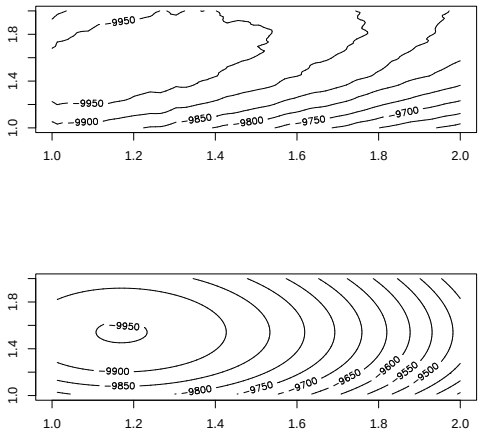

FIG. 2 – optimizations under nonparametric and semiparametric assumptions

Data and Estimation

Suppose that n item are observed. The ith item is observed under the vector of covariates $\mathbf{x}^{(i)}$ at m_i ordered moments t_{ij}. The values of the degradation levels $Z(t_{ij})$ are supposed to be measured without errors. We have thus for the ith item the successives values of increments of degradation, the value $\mathbf{x}^{(i)}$ describing the environment, the indicator of censorship $\delta_i = 1_{\{T_i \leq C_i\}}$.

The estimates are based on maximum likelihood or pseudo-maximum likelihood estimates. Numerical optimization is needed to achieve this estimation. We will compare through simulations the efficiency of the method for two different assumptions on the mean degradation function m. We will discuss also the choice of numerical algorithms for the optimization of the likelihood for the two models. We will consider

- **Model 1** $EZ(t) = m(t, \theta), \quad t > 0$, where m is a known function and $\theta \in I\!R^p$ is a parameter to be estimated.
- **Model 2** $EZ(t) = m(t), \quad t > 0$, where m is a unknown function to be estimated.

In both model, we supposed that :$\lambda(t) = \lambda(Z(t)) = \alpha Z(t)^p, \quad \alpha > 0, \quad p \geq 1$ where α and p are two unknown coefficients to be estimated. X is a discrete random variable describing p different environments. We have chosen p=3.

Different subsets of simulations have been carried out in order to measure the influence of the number of item, the number of degradation data for each item, the type of assumption (parametric versus nonparametric) and the issue of lack-of-fit in the parametric model.

References

- Bagdonavicius, V., Nikulin, M. (2000). Estimation in degradation Models with Explanatory Variables, Lifetine Data Analysis, **7**, 85-103.
- Bagdonavicius, V., Nikulin, M. (2001). Accelerated Life Models : Modeling and Statistical Analysis, Chapman & Hall / CRC, London.
- Bagdonavicius, V., Gerville-Réache, L., Nikulin, M., Nikoulina, V. (2000). Expériences accélérées : analyse statistique du modèle standard de vie accélérée, Revue de Statistique appliquée **XLVIII**, 3.
- Meeker, W.Q., Escobar, L. (1998). Statistical Methods for Reliability Data, John Wiley and Sons, N.Y.
- Nelson, W., (1990). Accelerated Testing : Statistical Models, Test Plan and Data Analysis, J.Wiley and Sons, N.Y.
- Singpurwalla , N.D. (1995). Survival in dynamic environnements, Statistical Science, 1, 10, 86-103.

.

Modèle Linéaire Généralisé et Directions Révélatrices

Michel DELECROIX
CREST-ENSAI
Campus de Ker-Lann
35170 Bruz
e-mail : delecroi@ensai.fr

Résumé

Les modèles linéaires généralisés sont d'utilisation plus que courante, en Actuariat par exemple, sans que soit jamais réellement vérifiée l'adéquation des hypothèses qu'ils présupposent aux jeux de données étudiés. On rappelle ici pourquoi une seule de ces hypothèses est en fait réellement nécessaire aux études statistiques menées, à condition d'utiliser une méthodologie appropriée, que l'on décrit. On s'interroge ensuite sur le sens de cette hypothèse et les possibilités de l'étendre. On propose enfin une méthode nouvelle pour la tester.

Le modèle de régression appelé "modèle linéaire généralisé", tel qu'il est défini dans le livre de Mac Cullagh et Nelder (1989), apparaît être l'instrument privilégié de nombreux praticiens de la Statistique. Pour citer un exemple précis, il reste l'outil unique (via "GENMOD" de SAS) de la mise à jour, à partir de jeux de données (les dossiers clients), de la valeur des primes d'assurance automobile, et plus généralement de toutes les analyses statistiques d'intérêt (analyse de la "sinistralité" des clients, prévision du nombre d'accidents par assuré, coût des sinistres) effectuées sur ces dossiers par les chargés d'étude des compagnies d'assurance, et leurs stagiaires étudiants.

Le problème posé à travers ces diverses situations peut aisément se schématiser : Pour chacun des n individus dont le contrat est déjà en portefeuille depuis un temps suffisant, on dispose de l'observation effective d'une variable d'intérêt, que nous appelerons Y et supposerons réelle, et d'un vecteur de d variables explicatives X. On est ainsi confronté à un problème d'estimation de l'espérance de la valeur de Y conditionnellement aux valeurs de X. Nous appelerons m cette espérance conditionnelle, notée classiquement

$$m(x) = E(Y|X = x).$$

La technologie statistique universellement utilisée, le "modèle linéaire généralisé", suppose vérifiés un certain nombre de postulats relatifs à la loi conjointe des variables Y et X. Il semble que nombre de statisticiens utilisant cette technologie, via un logiciel "presse-bouton", ne se rappellent pas toujours qu'en fait elle est exactement adaptée si on peut admettre que :

1 La loi de Y, conditionnellement aux valeurs de X, est une des lois du "modèle exponentiel". L'influence de la valeur x prise par X s'exprime à travers un (seul usuellement) des paramètres caractéristiques de cette loi, qu'on peut définir comme étant son espérance, modulo une éventuelle transformation des notations de Mac Cullagh et Nelder (1989).

2 Cette espérance, soit $m(x)$ selon nos notations, s'exprime comme une fonction d'une combinaison linéaire des composantes X_i de X, que nous noterons $\beta.X$, le point représentant le produit scalaire usuel. Nous appelerons "fonction de lien" cette fonction, notée h, introduisant, là aussi, une légère distorsion avec le vocabulaire usuel en la matière. Cette seconde hypothèse affirme donc l'existence de la fonction h et du d-uple β tels que, pour toute valeur x, on ait :

$$m(x) = h(\beta.x).$$

La fonction h est supposée connue, choisie par le statisticien. Le modèle est donc paramétrique, de paramètre β.

Ayant admis cette hypothèse générale, le statisticien doit finalement choisir la loi conditionnelle exacte (Poisson, Bernoulli, Gamma, Normale, ...) et la fonction h qu'il va utiliser. Le premier choix se fait usuellement à partir de la nature des valeurs prises par Y, le second est souvent réalisé par le logiciel lui-même (option par défaut), de telle sorte que les calculs à mener (résoudre les équations de vraisemblance) soient simplifiés (choix des "fonctions de lien canoniques"). Le modèle obtenu étant totalement paramétrique, les techniques "standard" du maximum de vraisemblance livrent un estimateur de β donc de $m(x)$.

Une question fondamentale en la matière est évidemment celle de l'adéquation du modèle aux données utilisées. Il est bien facile de trouver des variables continues non gaussiennes, à valeurs entières positives ne suivant pas la loi de Poisson, etc ..., et il est en général extrêmement difficile de justifier a priori le choix de telle ou telle fonction de lien, une justification a posteriori pouvant être de comparer empiriquement les résultats que donnent des choix divers. Finalement, il n'existe pas de test d'adéquation global du modèle aux données. Si l'on veut éviter la classique pirouette résumée par la formule : "Tous les modèles sont faux, quelques-uns sont utiles", il apparaît légitime de s'interroger sur les procédés alternatifs d'estimation de $m(x)$, sous des hypothèses moins contraignantes et testables.

Une solution possible est l'approche "non-paramétrique", qui supprime toute hypothèse particulière sur la loi du couple (X, Y). Cette approche séduisante se révèle cependant inadaptée aux jeux de données dont on dispose usuellement, puisqu'exigeant par exemple, que toutes les variables X_i soient continues, ainsi qu'un trés grand nombre de données ("fléau de la dimension"). Elle souffre aussi beaucoup d'un manque d'interprétabilité des résultats obtenus (l'estimation de m se fait point par point) et de la difficulté du choix d'un paramètre de lissage, souvent très délicat.

Si l'on interprète la remarque précédente comme le fait que supprimer totalement les hypothèses du MLG empêche de mener l'analyse statistique recherchée, on peut alors s'interroger sur les solutions existantes quand on ne garde qu'une des deux hypothèses du modèle. Garder la première sans la seconde mène aux méthodes de "vraisemblance locale" détaillées par exemple dans le livre de Fan et Gijbels (1996). Celles-ci souffrent finalement des mêmes défauts que les méthodes non-paramétriques évoquées ci-dessus. Reste alors la solution d'estimer la régression m sous la seconde hypothèse seulement, c'est-à-dire sous la seule hypothèse d'existence d'une "**direction révélatrice**" unique, sans que soit même spécifiée la fonction de lien.

Pour montrer la pertinence de cette idée, on montrera d'abord que l'hypothèse peut se généraliser "naturellement" en celle de l'existence de plusieurs directions révélatrices. On passera ensuite en revue les méthodes d'estimation du paramètre et de la fonction de régression, et de test de l'existence et du nombre des directions révélatrices. on proposera enfin une statistique inédite qui assure au test de la seconde hypothèse vue ci-dessus une puissance correcte sous une suite d'alternatives locales.

Références

- Mac Cullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models*, Chapman et Hall, London.
- Fan, J. and Gijbels, I. (1996) *Local Polynomial Modelling and Its Applications*, Chapman et Hall, London.

.

Etude asymptotique de l'Analyse Canonique : de l'approche matricielle et analytique à l'approche opératorielle et tensorielle.

Jeanne FINE
Laboratoire Statistique et Probabilités
Université Paul Sabatier
31062 Toulouse Cedex
e-mail : fine@cict.fr

Résumé

L'étude asymptotique de l'Analyse Canonique (Anderson, 1999, Fine, 2000) donne l'occasion dans cet exposé de montrer l'intérêt de l'approche opératorielle et tensorielle de la statistique asymptotique multidimensionnelle par rapport à l'approche classique, matricielle et analytique. On insiste sur le cadre de l'analyse (Dauxois et Pousse, 1976), le modèle d'échantillonnage (Dauxois, Fine et Pousse, 1979) et les différents outils mathématiques (Fine, 1987, Dauxois, Romain et Viguier, 1994) qui permettent de résoudre les problèmes posés dans ce type d'étude.

1. Approche classique, matricielle et analytique

1.1. Définition de l'Analyse Canonique

Soit X et Y deux vecteurs aléatoires, de dimension p et q respectivement ($p \leq q$), définis sur un espace probabilisé $(\Omega, \mathcal{A}, P)$, supposés centrés et admettant des moments d'ordre 2, V_X (resp. V_Y) la matrice de covariance supposée inversible de X (resp. de Y). On note H_X (resp. H_Y) l'espace vectoriel des v.a.r. combinaisons linéaires des composantes de X (resp. Y).

L'Analyse Canonique (A.C.) du couple (X, Y) a pour objectif de mesurer les liens entre X et Y et peut être définie comme la recherche d'un premier couple (f_1, g_1) de v.a.r. de H_X et H_Y, de variance 1 et de corrélation ρ'_1 maximale puis, de façon itérée, pour $j = 2, ..., r$, $(r \leq p)$, d'un couple (f_j, g_j) de v.a.r. de H_X et

H_Y, de variance 1, non corrélées avec les $(f_k)_{k<j}$ et les $(g_k)_{k<j}$ et de corrélation ρ'_j maximale. Le couple (f_j, g_j) est le $j^{ème}$ couple de *variables canoniques* et le réel ρ'_j de [0,1] est le $j^{ème}$ *coefficient de corrélation canonique.*

Soit V_{XY} (resp. V_{YX}) la matrice des covariances des composantes de X (resp. Y) avec celles de Y (resp. X). On pose $R_X = V_X^{-\frac{1}{2}} V_{XY} V_Y^{-1} V_{YX} V_X^{-\frac{1}{2}}$ (resp. $R_Y = V_Y^{-\frac{1}{2}} V_{YX} V_X^{-1} V_{XY} V_Y^{-\frac{1}{2}}$), alors R_X et R_Y ont mêmes valeurs propres (v.p.) non nulles et, si r désigne le rang commun de R_X et R_Y, $(\rho'^2_j)_{j=1,...,r}$ est la "suite pleine décroissante" des v.p. non nulles de R_X et R_Y (c'est-à-dire, la suite décroissante des v.p. répétées selon l'ordre de multiplicité de ces v.p.). On pose : $\lambda'_j = \rho'^2_j$, pour $j = 1, ..., r$, et, sauf dans le cas particulier $r = p = q$, on pose $\lambda'_j = \rho'_j = 0$ pour $j > r$. Pour $j = r+1, ..., p$ (resp. $j = r+1, ..., q$) on définit f_j (resp. g_j) comme v.a.r. de H_X (resp. H_Y), de variance 1, non corrélée avec les $(f_k)_{k<j}$ et $(g_k)_{k<j}$.

L'Analyse Canonique du couple (X, Y) est alors le triplet :

$$((\rho'_j)_{j=1,...,r+1}, (f_j)_{j=1,...,p}, (g_j)_{j=1,...,q}). \tag{9}$$

L'A.C. de (X, Y) ne dépend que des espaces H_X et H_Y, espaces qui sont également engendrés respectivement par les composantes de $X' := V_X^{-\frac{1}{2}} X$ et de $Y' := V_Y^{-\frac{1}{2}} Y$. On montre que si $(u_j)_{j=1,...,p}$ (resp. $(v_j)_{j=1,...,q}$) désigne une base de vecteurs propres unitaires de R_X (resp. R_Y) associés à $(\lambda'_j)_{j=1,...,p}$ (resp. $(\lambda'_j)_{j=1,...,q}$) on peut obtenir les variables canoniques f_j et g_j comme combinaisons linéaires des composantes de X' et Y' en utilisant les composantes de u_j et v_j comme coefficients, c'est-à-dire, en posant : $f_j = \langle u_j, X' \rangle_p$ et $g_j = \langle v_j, Y' \rangle_q$ où $\langle ., . \rangle_p$ et $\langle ., . \rangle_q$ désignent les produits scalaires usuels de $\mathbb{R}^p$ et $\mathbb{R}^q$ respectivement.

Le triplet (1) n'est pas unique dans la mesure où chaque variable canonique associée à une v.p. simple peut être remplacée par son opposée et que l'ensemble des variables canoniques associées à une v.p. multiple peut être remplacé par un autre ensemble en fonction du choix des bases des espaces propres de R_X et R_Y associés à la v.p. multiple.

1.2. Définition de l'Analyse Canonique d'échantillonnage

Soit $(X_l, Y_l)_{l=1,...,n}$ un échantillon de taille n i.i.d. comme (X, Y). On indexe par n les éléments définis précédemment et calculés sur l'échantillon :

$$\mu_X^n, \mu_Y^n, V_X^n, V_Y^n, V_{XY}^n, R_X^n, R_Y^n.$$

Soit $(\lambda'^n_j)_{j=1,...,p}$ la suite pleine décroissante des p v.p. de R_X^n (et des p plus grandes v.p. de R_Y^n, les autres (si $q > p$) étant nulles), $(u_j^n, v_j^n)_{j=1,...,p}$ une suite de vecteurs propres unitaires associés de R_X^n et de R_Y^n et $(f_j^n, g_j^n)_{j=1,...,p}$ la suite des

variables canoniques, vecteurs de $\mathbb{R}^n$, obtenues par :

$$\forall l \in \{1, ..., n\} \quad (f_j^n)_l = \langle u_j^n, (V_X^n)^{-\frac{1}{2}}(X_l - \mu_X^n)\rangle_p \quad \text{et} \quad (g_j^n)_l = \langle v_j^n, (V_Y^n)^{-\frac{1}{2}}(Y_l - \mu_Y^n)\rangle_q.$$

Le cas échéant ($q > p$), on pose $\lambda'^n_{p+1} = 0$, on complète $(v_j^n)_{j=1,...,p}$ en une base orthonormée de $\mathbb{R}^q$ de vecteurs propres $(v_j^n)_{j=1,...,q}$ de R_Y^n et on définit les variables canoniques associées.

Enfin, pour tout j de $\{1, ..., p+1\}$ on pose $\rho'^n_j = \sqrt{\lambda'^n_j}$, l'A.C. d'échantillonnage de (X, Y) est alors :

$$((\rho'^n_j)_{j=1,...,p+1}, (f_j^n)_{j=1,...,p}, (g_j^n)_{j=1,...,q}). \tag{10}$$

1.3. Étude asymptotique

On suppose désormais que le couple (X, Y) admet des moments d'ordre 4. Réaliser une étude asymptotique de l'A.C. consiste à établir la convergence p.s. des différents éléments de l'A.C. d'échantillonnage vers les éléments correspondants de l'A.C. théorique et la convergence en loi des éléments standardisés correspondants.

Les difficultés sont nombreuses : les variables canoniques sont des variables aléatoires estimées ("prédites") par des vecteurs de $\mathbb{R}^n$, espace dont la dimension augmente avec la taille de l'échantillon. Il est alors d'usage de restreindre l'étude asymptotique aux vecteurs propres $(u_j^n)_{j=1,...,p}$ (resp. $(v_j^n)_{j=1,...,q}$) de R_X^n (resp. R_Y^n), appelés *vecteurs canoniques* et aux vecteurs de $\mathbb{R}^p$ (resp. $\mathbb{R}^q$) définis par : $x_j^n = (V_X^n)^{-\frac{1}{2}}u_j^n$ (resp. $y_j^n = (V_Y^n)^{-\frac{1}{2}}v_j^n$), appelés *facteurs canoniques*, qui permettent d'obtenir les variables canoniques de façon directe :

$$\forall l \in \{1, ..., n\} \quad (f_j^n)_l = \langle x_j^n, X_l - \mu_X^n\rangle_p \quad \text{et} \quad (g_j^n)_l = \langle y_j^n, Y_l - \mu_Y^n\rangle_q.$$

Le cas de valeurs propres multiples est difficile à traiter car les vecteurs propres associés ne sont pas définis de façon unique. Il est alors d'usage de restreindre l'étude asymptotique au cas où toutes les v.p. non nulles sont simples. L'unicité est alors vérifiée en choisissant systématiquement le vecteur propre unitaire (parmi les deux) dont la première coordonnée non nulle par rapport à la base canonique est strictement positive.

Comme pour toutes les analyses multidimensionnelles, les matrices aléatoires d'échantillonnage, V_X^n, V_{XY}^n, R_X^n, ... ont pour matrice de covariance des "super-matrices" (c'est-à-dire, des matrices de matrices). Des outils, comme l'opérateur "vec" de vectorisation d'une matrice, ont alors été introduits afin de manipuler ces super-matrices, la difficulté essentielle résidant dans la nécessité de préciser l'ordre des éléments des lignes et des colonnes.

Par ailleurs, on sait que la suite $(\sqrt{n}(V_X^n - V_X))$ converge en loi vers une gaussienne centrée dont la super-matrice de covariance a une forme connue dans certains cas particuliers : X gaussienne ou elliptique par exemple.

Dans le cadre de l'A.C., il s'agit d'étudier la convergence de la suite $(\sqrt{n}(V_Z^n - V_Z))$ avec $Z = (X, Y)$ et donc $V_Z = \begin{pmatrix} V_X & V_{XY} \\ V_{YX} & V_Y \end{pmatrix}$ et $V_Z^n = \begin{pmatrix} V_X^n & V_{XY}^n \\ V_{YX}^n & V_Y^n \end{pmatrix}$,

puis la convergence de la suite $(\sqrt{n}(R_Z^n - R_Z))$ avec $R_Z = (R_X, R_Y)$ et $R_Z^n = (R_X^n, R_Y^n)$, avant d'étudier la convergence des suites d'éléments propres de R_Z^n puis des suites des différents éléments de l'A.C. d'échantillonnage.

Ce n'est qu'en 1999 qu'Anderson publie les résultats de cette étude asymptotique pour un couple (X, Y) gaussien et dans le cas où toutes les valeurs propres non nulles sont simples. Les composantes des facteurs canoniques et les coefficients de corrélations canoniques de l'A.C. théorique (resp. d'échantillonnage) sont des fonctions différentiables de V_Z (resp. de V_Z^n). Les résultats sont alors obtenus à partir de développements de Taylor. L'approche classique peut ainsi être qualifiée de matricielle et analytique.

Afin de simplifier les calculs on fait un changement de variables de (X, Y) à (X', Y'), ce qui est équivalent à effectuer un changement de bases dans $\mathbb{R}^p$ et $\mathbb{R}^q$. On a alors : $V_{X'} = I_p$, (resp. $V_{Y'} = I_q$) et $R_X = R_{X'} = V_{X'Y'}V_{Y'X'}$ (resp. $R_Y = R_{Y'} = V_{Y'X'}V_{X'Y'}$).

2. Approche opératorielle et tensorielle

2.1. Introduction et historique

Les difficultés qui viennent d'être soulevées sont essentiellement dues au fait que l'outil matriciel n'est pas adapté. Travailler directement sur les opérateurs linéaires d'espaces euclidiens évite les problèmes d'indexation et peut être facilement généralisé à un cadre hilbertien. De plus, au lieu d'étudier les vecteurs propres associés à des v.p. simples, il est possible d'étudier les projecteurs propres associés à des v.p. multiples. Eaton (1983) prône également l'approche opératorielle de la statistique multidimensionnelle.

Dès 1976, Dauxois et Pousse élargissent la définition de l'Analyse en Composantes Principales (A.C.P.) d'un vecteur aléatoire de $\mathbb{R}^p$ à celle d'une v.a. hilbertienne (v.a.h.) et même d'une fonction aléatoire hilbertienne (f.a.h.), c'est-à-dire, d'une v.a.h. dépendant d'un paramètre afin de prendre en compte les données chronologiques ou spatiales. Ils reprennent l'ensemble des méthodes factorielles (A.C.P., A.C., Analyse Factorielle des Correspondances, Analyse Factorielle Discriminante, ...) dans un cadre opératoriel et probabiliste, ce qui les amène, entre autres, à définir des analyses non-linéaires.

La première étude asymptotique dans ce cadre est réalisée par Romain pour l'A.C.P. d'une f.a.h. en 1979 (cf. aussi Dauxois, Pousse et Romain, 1982), étude complétée par Arconte en 1980 qui aborde aussi l'étude asymptotique de l'A.C. mais tous les outils ne sont pas encore disponibles pour poursuivre l'étude. En 1994, Dauxois, Romain et Viguier proposent d'utiliser d'autres produits tensoriels et établissent un dictionnaire entre les écritures matricielles et les écritures opératorielles. Ce travail permet de comparer plus aisément des résultats communs qui seraient obtenus dans l'un et l'autre cadre, mais il permet aussi d'obtenir plus commodément des résultats complexes ; les écritures par rapport aux bases de vecteurs propres ne sont données qu'après avoir établi les résultats sous forme concise sur les opérateurs.

Ces nouveaux outils permettent de réaliser l'étude asymptotique de l'A.C. dans toute sa généralité (Fine, 2000), c'est-à-dire, sans hypothèse sur la distribution de probabilité du couple de v.a. euclidiennes dont on fait l'analyse, dans le cas général de v.p. éventuellement multiples et sans écarter l'étude du comportement asymptotique des variables canoniques (éléments aléatoires de l'A.C.). Notre approche peut donc être qualifiée d'opératorielle et tensorielle. Nous précisons ci-après les différentes étapes de l'étude asymptotique de l'A.C. et les outils mathématiques utilisés et donnons quelques résultats.

2.2. Les différentes étapes de l'étude asymptotique de l'A.C., les outils, les résultats

1) A.C. théorique

Il s'agit tout d'abord de définir l'A.C. d'un couple de v.a. euclidiennes (v.a.e.) (A.C. théorique). On reprend les notations de l'approche classique en remplaçant $(\mathbb{R}^p, \langle ., . \rangle_p)$ (resp. $(\mathbb{R}^q, \langle ., . \rangle_q)$ par un espace euclidien $(\mathcal{X}, \langle ., . \rangle_{\mathcal{X}})$ (resp. $(\mathcal{Y}, \langle ., . \rangle_{\mathcal{Y}})$) de dimension p (resp. q).

Soit $L^2(P)$ l'espace de Hilbert des v.a.r. d'ordre 2 définies sur $(\Omega, \mathcal{A}, P)$, muni du produit scalaire qui à (f, g) associe $\mathbb{E}(fg)$.

L'opérateur Φ_X de $\mathcal{X}$ dans $L^2(P)$ qui à x associe $\langle x, X \rangle_{\mathcal{X}}$ joue un rôle essentiel dans l'approche opératorielle de la statistique multidimensionnelle. L'espérance mathématique de X est l'unique élément de $\mathcal{X}$ (théorème de Riesz), noté $\mathbb{E}(X)$ ou μ_X, vérifiant : $\forall x \in \mathcal{X}, \langle x, \mathbb{E}(X) \rangle_{\mathcal{X}} = \mathbb{E}(\langle x, X \rangle_{\mathcal{X}})$.

Pour tout $(x, y) \in \mathcal{X} \times \mathcal{Y}$, on note $x \otimes y$ l'opérateur de $\mathcal{X}$ dans $\mathcal{Y}$ qui à x' associe $\langle x', x \rangle_{\mathcal{X}}\ y$, élément de l'espace de Hilbert $\sigma_2(\mathcal{X}, \mathcal{Y})$ des opérateurs de $\mathcal{X}$ dans $\mathcal{Y}$ muni du produit scalaire : $\langle A, B \rangle_2 = tr(AB^*)$. Grâce au théorème de Riesz, on peut alors définir les opérateurs de covariance V_X de X, V_Y de Y, et les opérateurs de covariance croisée V_{XY} et V_{YX} de X et Y : $V_X = \mathbb{E}((X - \mu_X) \otimes (X - \mu_X))$, ...

Sans perte de généralité on suppose désormais que les v.a. X et Y sont centrées

et d'ordre 4. L'adjoint Φ_X^* de Φ_X est l'opérateur de $L^2(P)$ dans $\mathcal{X}$ qui à f associe $\mathbb{E}(fX)$ et on a donc : $\Phi_X^* \circ \Phi_X = V_X$, $\Phi_X^* \circ \Phi_Y = V_{XY}$, ... opérateurs mis en relations dans le schéma de dualité suivant :

$$\begin{array}{ccccc} \mathcal{X} & \overset{\Phi_X^*}{\longleftarrow} & L^2(P) & \overset{\Phi_Y^*}{\longrightarrow} & \mathcal{Y} \\ V_X^{-1} \downarrow\uparrow V_X & & \uparrow I & & V_Y \uparrow\downarrow V_Y^{-1} \\ \mathcal{X} & \overset{\longrightarrow}{\Phi_X} & L^2(P) & \overset{\longleftarrow}{\Phi_Y} & \mathcal{Y} \end{array}$$

Les opérateurs R_X et R_Y sont définis comme précédemment (les lois de composition $\circ$ sont supprimées pour alléger les écritures), ainsi que l'A.C. du couple (X, Y).

Comme dans l'approche classique, afin de simplifier les calculs, nous changeons de produit scalaire sur $\mathcal{X}$ (resp. $\mathcal{Y}$) de façon à ce que l'opérateur de covariance de X (resp. de Y) soit l'identité de $\mathcal{X}$ (resp. $\mathcal{Y}$). L'opérateur R_X (resp. R_Y) s'écrit alors : $R_X = V_{XY}V_{YX}$ (resp. $R_Y = V_{YX}V_{XY}$).

2) Modèle d'échantillonnage et A.C. d'échantillonnage

On utilise un modèle d'échantillonnage (Dauxois, Fine et Pousse, 1979) faisant le lien entre l'échantillon utilisé en Analyse des Données et l'échantillon statistique i.i.d. de la Statistique classique. L'échantillon $(X_l, Y_l)_{l\in\mathbb{N}^*}$ i.i.d. comme (X, Y) est construit à partir d'un élément ω de $\Omega^{\mathbb{N}^*}$ en posant, pour tout l de $\mathbb{N}^*$ (π_l désignant la $l^{ème}$ projection de $\Omega^{\mathbb{N}^*}$ sur Ω) : $X_l = X \circ \pi_l$ et $Y_l = Y \circ \pi_l$ c'est-à-dire $X_l(\omega) = X(\omega_l)$ et $Y_l(\omega) = Y(\omega_l)$.

On munit alors $L^2(P)$ du produit scalaire (aléatoire puisqu'il dépend de ω) :

$$\forall (f, g) \in L^2(P) \otimes L^2(P), \quad \mathbb{E}_n(fg) = \frac{1}{n}\sum_{l=1}^{n} f(\omega_l)g(\omega_l).$$

On a alors : $\mathbb{E}_n(X) = \mu_X^n$, $\Phi_X^n = \langle ., X - \mu_X^n\rangle_{\mathcal{X}}$, $V_X^n = \frac{1}{n}\sum_{l=1}^{n}(X_l - \mu_X^n) \otimes (X_l - \mu_X^n)$, ...

Le schéma de dualité de l'A.C. d'échantillonnage est celui de l'A.C. théorique en remplaçant $L^2(P)$ par $(L^2(P), \mathbb{E}_n)$ et en indexant par n les opérateurs. Les opérateurs R_X^n et R_Y^n et l'A.C. d'échantillonnage sont définis comme précédemment.

3) Convergences p.s. et en loi des opérateurs d'échantillonnage

Les théorèmes limites dans les espaces euclidiens ou hilbertiens permettent d'obtenir les convergences p.s. et en loi de la suite des opérateurs de covariance d'échantillonnage sans aucune hypothèse sur la distribution des v.a. autre que l'existence de moments d'ordre 4. Pour l'A.C., on obtient en particulier (on rappelle que l'on a posé $Z = (X, Y)$) :

$$W_Z^n := \sqrt{n}(V_Z^n - V_Z) \quad \overset{\mathcal{L}oi}{\longrightarrow} \quad W_Z \sim N(0\ ;\ \mathbb{K}_Z),$$

où $\mathbb{K}_Z$ est l'opérateur de covariance de $(Z-\mu_Z)\otimes(Z-\mu_Z)$.

Concernant les opérateurs d'échantillonnage $R_Z^n=(R_X^n,R_Y^n)$ de $\sigma_2(\mathcal{Z})$ (avec $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$), la convergence p.s. vient du fait que l'on peut écrire R_X^n et R_Y^n comme fonctions continues de V_Z^n ; on pose $U_Z^n=\sqrt{n}(R_Z^n-R_Z)$ $(:=(U_X^n,U_Y^n))$ et on écrit $U_X^n=\Psi_X^n(W_Z^n)$ (resp. $U_Y^n=\Psi_Y^n(W_Z^n)$) où (Ψ_X^n) (resp. (Ψ_Y^n)) est une suite d'opérateurs aléatoires de $\sigma_2(\mathcal{Z})$ dans $\sigma_2(\mathcal{X})$ (resp. $\sigma_2(\mathcal{Y})$) convergeant p.s. vers Ψ_X (resp. Ψ_Y).

On peut alors en déduire la convergence en loi de (U_X^n) (resp. (U_Y^n)) vers $U_X=\Psi_X(W_Z)$ (resp. $U_Y=\Psi_Y(W_Z)$), gaussienne centrée d'opérateur de covariance $\mathbb{L}_X=\Psi_X\circ\mathbb{K}_Z\circ\Psi_X^*$ (resp. $\mathbb{L}_Y=\Psi_Y\circ\mathbb{K}_Z\circ\Psi_Y^*$). Le théorème utilisé, qu'il est facile de montrer à partir de résultats classiques dans les espaces métriques (Billingsley, 1968) a été attribué à tort dans nos travaux précédents à Rubin (qui envisage des fonctions non linéaires mais également non aléatoires).

On obtient pour U_X (et l'analogue pour U_Y en échangeant les rôles de X et Y) :

$$U_X=-\frac{1}{2}(W_XR_X+R_XW_X)+W_{XY}V_{YX}+V_{XY}W_{YX}-V_{XY}W_YV_{YX},$$

4) Convergences p.s. et en loi des éléments propres et des éléments de l'A.C.

Quelle que soit la méthode "factorielle" en jeu, c'est-à-dire, une analyse ou un modèle dont les éléments sont obtenus à partir d'une décomposition spectrale, les résultats concernant les éléments propres (v.p. simples ou multiples, projecteurs propres, vecteurs propres associés à des valeurs propres simples, ...) sont obtenus aisément grâce à la théorie des perturbations d'opérateurs linéaires (Kato, 1980). En Fine (1987), la théorie a été adaptée à des perturbations bornées ce qui, grâce à la loi du logarithme itéré, permet de l'utiliser dans le cadre des études asymptotiques par échantillonnage pour obtenir des développements presque sûrs des éléments propres d'une suite d'opérateurs aléatoires autoadjoints positifs.

On pourra aussi consulter Dossou-Gbete et Pousse, 1991, pour les résultats limites mais, pour la convergence en loi de certains éléments de l'A.C., les résultats limites ne sont pas suffisants et il faudra revenir aux développements des perturbations.

Prenons l'exemple des facteurs canoniques associés à une valeur propre simple λ_i ; on a alors : $x_i=u_i$ car $V_X=I_{\mathcal{X}}$ et $x_i^n=(V_X^n)^{-\frac{1}{2}}u_i^n$ d'où :

$$\sqrt{n}(x_i^n-x_i)=-(V_X^n)^{-\frac{1}{2}}((V_X^n)^{\frac{1}{2}}+I_{\mathcal{X}})^{-1}[\sqrt{n}(V_X^n-I_{\mathcal{X}})]u_i^n+[\sqrt{n}(u_i^n-u_i)].$$

On sait que $(\sqrt{n}(V_X^n-I_{\mathcal{X}}))$ converge en loi vers W_X et $(\sqrt{n}(u_i^n-u_i))$ vers $S_{X_i}U_Xx_i$ (avec $S_{X_i}=(R_X-\lambda_iI_{\mathcal{X}})^-$) mais c'est grâce aux développements des perturbations

que l'on pourra établir que la suite $(\sqrt{n}(x_i^n - x_i))$ converge en loi vers $\frac{1}{2}W_X x_i + S_{X_i} U_X x_i$.

5) Explicitation des opérateurs de covariance asymptotique dans le cas général puis dans le cas elliptique

Nous avons déjà vu que l'opérateur de covariance asymptotique de $(\sqrt{n}(R_X^n - R_X))$ est $\mathbb{L}_X = \Psi_X \circ \mathbb{K}_Z \circ \Psi_X^*$, où $\mathbb{K}_Z$ est l'opérateur de covariance asymptotique de $(\sqrt{n}(V_Z^n - V_Z))$ et où l'opérateur Ψ_X de $\sigma_2(\mathcal{Z})$ dans $\sigma_2(\mathcal{X})$ peut être explicité. Toutes les limites en loi des suites d'éléments propres ou d'éléments canoniques sont des gaussiennes centrées (ou des fonctions de gaussiennes centrées) dont on peut écrire l'opérateur de covariance en fonction de $\mathbb{K}_Z$ de façon analogue.

Tous ces opérateurs de covariance asymptotique s'explicitent dans le cas où Z suit une loi elliptique d'espérance μ_Z, d'opérateur de covariance V_Z et de kurtosis κ (paramètre réel, qui, lorsqu'il est nul, conduit à une loi $N(\mu_Z, V_Z)$). On sait alors que $\mathbb{K}_Z$ est l'opérateur de $\sigma_2(\mathcal{Z})$ dans lui-même qui à T associe :

$$\mathbb{K}_Z(T) = (1+\kappa)V_Z(T + T^*)V_Z + \kappa\langle V_Z, T\rangle_2 V_Z.$$

C'est à ce niveau que les explicitations nécessitent des outils algébriques supplémentaires. Le produit tensoriel dans des espaces de type σ_2 sera noté $\tilde{\otimes}$. Par exemple : $\forall (A,B) \in \sigma_2(\mathcal{Z}) \times \sigma_2(\mathcal{Z}), \quad \forall T \in \sigma_2(\mathcal{Z}), \quad A\tilde{\otimes}B(T) = \langle T, A\rangle_2 B$. On définit aussi le produit $\overset{\ell}{\otimes}$ dans les espaces de type σ_2. Par exemple : $\forall (A,B) \in \sigma_2(\mathcal{Z}) \times \sigma_2(\mathcal{Z}), \quad \forall T \in \sigma_2(\mathcal{Z}), \quad A \overset{\ell}{\otimes} B(T) = BTA^*$. On définit l'opérateur de commutation $\mathcal{C}$ qui à un opérateur T associe son adjoint T^*, enfin, on remplace le produit $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ par la somme hilbertienne $\mathcal{Z} = \mathcal{X} \oplus \mathcal{Y}$ ce qui permet de plonger tous les opérateurs dans $\sigma_2(\mathcal{Z})$ et simplifie les écritures. Le projecteur P_X de $\sigma_2(\mathcal{Z})$ sur $\sigma_2(\mathcal{X})$ est à présent un opérateur autoadjoint de $\sigma_2(\mathcal{Z})$ dans lui-même, ...

L'opérateur $\mathbb{K}_Z$ de $\sigma_2(\mathcal{Z})$ dans lui-même peut s'écrire :

$$\mathbb{K}_Z = (1+\kappa)V_Z \overset{\ell}{\otimes} V_Z(I + \mathcal{C}) + \kappa V_Z \tilde{\otimes} V_Z.$$

Soit $(x_i)_{i=1,\ldots,p}$ une base orthonormée de facteurs canoniques de $\mathcal{X}$, alors $(x_i \otimes x_j)_{i,j=1,\ldots,p}$ est une base orthonormée de $\sigma_2(\mathcal{X})$ et $((x_i \otimes x_j)\tilde{\otimes}((x_k \otimes x_l))_{i,j,k,l=1,\ldots,p}$ est une base de $\sigma_2(\sigma_2(\mathcal{X}))$.

Après avoir réalisé les calculs de façon synthétique, il est aisé de décomposer les opérateurs par rapport à ce type de bases. On obtient, par exemple, pour l'opérateur de covariance asymptotique de $(\sqrt{n}(R_X^n - R_X))$:

$$\mathbb{L}_{XX} = (1+\kappa)(I+\mathcal{C})[-\frac{3}{4}R_X^2 \overset{\ell}{\otimes} I_{\mathcal{X}} + R_X^2 \overset{\ell}{\otimes} R_X + R_X \overset{\ell}{\otimes} I_{\mathcal{X}} - \frac{5}{4}R_X \overset{\ell}{\otimes} R_X](I+\mathcal{C})$$

et, par rapport aux bases de facteurs canoniques (on rappelle que $(\lambda'_j)_{j=1,\ldots,p}$ est la suite pleine décroissante des v.p. de R_X) :

$$\begin{aligned}\mathbb{L}_{XX} &= \frac{1}{2}(1+\kappa)\sum_{j=1}^{p}\sum_{k=1}^{p}(-\frac{3}{4}\lambda'^2_j - \frac{3}{4}\lambda'^2_k + \lambda'^2_j\lambda'_k + \lambda'_j\lambda'^2_k + \lambda'_j + \lambda'_k - \frac{5}{2}\lambda'_j\lambda'_k) \\ &\quad (x_j \otimes x_k + x_k \otimes x_j)\tilde{\otimes}(x_j \otimes x_k + x_k \otimes x_j)\end{aligned}$$

Dans le cas gaussien et lorsque toutes les v.p. sont simples, il est possible de comparer nos résultats avec ceux d'Anderson. Nous sommes en désaccord sur deux d'entre eux.

6) Convergences p.s. et en loi des éléments aléatoires de l'A.C.

Le modèle d'échantillonnage utilisé (cf. § 2.1.2) permet de bien distinguer l'aléatoire qui vient du modèle (les variables canoniques, les espaces H_X et H_Y de $L^2(P)$ de l'A.C. théorique, ...) et l'aléatoire qui vient de l'échantillonnage. Il est alors possible d'obtenir les convergences p.s. et en loi des suites de variables canoniques.

7) Applications inférentielles et conclusion

Les résultats pour l'A.C. permettent d'aborder aisément les applications inférentielles (estimation par intervalle de confiance, test statistique, ...) impliquant les éléments d'une A.C., en particulier, les mesures de proximités entre deux ensembles de variables construites en fonction des coefficients de corrélation canonique. Cf. Dauxois et Nkiet, 2002, Anderson, 1999, ... En conclusion, nous espérons avoir montré que les différents outils mathématiques présentés ici pour l'étude asymptotique de l'A.C. sont très performants pour travailler en statistique asymptotique multidimensionnelle.

Références

- Anderson, T.W. (1999). Asymptotic Theory for Canonical Correlation Analysis. Journal of Multivariate Analysis, 70 1-29.
- Arconte, A. (1980). Étude asymptotique de l'analyse en composantes principales et de l'analyse canonique. Thèse de 3ème cycle, Université de Pau et des Pays de l'Adour.
- Billingsley, P. (1968). Convergence of probability measures. Wiley, New-York.
- Dauxois, J., Fine, J. et Pousse A. (1979). Échantillonnage en segmentation, étude de la convergence. Statistique et Analyse des Données, 3, 45-53.
- Dauxois, J. et Nkiet, G.M. (1997). Canonical Analysis of two Euclidean subspaces and its applications. Linear Algebra Appl., 264, 355-388.

- Dauxois, J. et Pousse, A. (1976). Les Analyses factorielles en calcul des Probabilités et en Statistique : essai d'étude synthétique. Thèse de Doctorat d'État, Université Paul Sabatier, Toulouse.
- Dauxois, J., Pousse, A. et Romain, Y. (1982). Asymptotic theory for the principal component analysis of a vector random function ; some applications to statistical inference. J. Multivariate Anal., 12, 136-154.
- Dauxois, J., Romain, Y. et Viguier, S. (1994). Tensor products and statistics. Linear Algebra Appl., 210, 59-88.
- Dossou-Gbete, S. et Pousse, A. (1991). Asymptotic study of eigenelements of a sequence of random self adjoint operators. Statistics, 22, 479-491.
- Eaton, M.L. (1983). Multivariate statistics. A vector space approach. Wiley, New-York.
- Fine, J. (1987). On the validity of the perturbation method in asymptotic theory. Statistics, 18, 401-414.
- Fine, J. (2000). Étude Asymptotique de l'Analyse Canonique. Pub. Inst. Stat. Univ. Paris, 44, 2-3, 21-72.
- Kato, T. (1980). Perturbation theory for linear operators. Springer-Verlag, New-York.
- Romain, Y. (1979). Étude Asymptotique des approximations par échantillonnage de l'analyse en composantes principales d'une fonction aléatoire. Quelques applications. Thèse 3ème cycle. Université Paul Sabatier. Toulouse.

Estimation du potentiel d'interaction de paires d'un processus de Gibbs sur des sphères concentriques dont on a observé plusieurs réalisations.

Jean-Michel BILLIOT et Michel GOULARD *
* Adresse pour correspondance :
INRA, Unité BIA, BP 27
31326 CASTANET-TOLOSAN Cedex, FRANCE
e-mail : goulard@toulouse.inra.fr

Résumé

Nous considérons la question de l'estimation du potentiel d'interaction de paire d'un processus de Gibbs. L'estimation de la fonction potentiel dans un cadre paramétrique et quand le support est le plan a été relativement bien étudié. Cette estimation est en général faite à partir d'une réalisation du processus. Nous voulons étudier le problème d'estimation quand plusieurs réalisations sont disponibles, situation qui peut arriver en agronomie ou en recherche médicale. Pour cela nous considérons un processus de Gibbs dans le cas où le support est une sphère, puis plusieurs sphères concentriques. Ce modèle semble adapté pour résumer l'insertion des racines autour d'un pivot de maïs, insertion sur laquelle nous disposons d'observations indépendantes de système racinaire de pied de maïs.

Nous développons une méthode d'essence fonctionnelle en utilisant une approximation par une série de Fourier. Dans le cas d'une série finie l'estimation est faite en minimisant un contraste approprié. Nous montrons que les estimateurs que nous définissons dans ce cadre sont consistants et asyptotiquement normaux. Selon une méthode assez classique ces résultats nous permettent de définir des tests pour des hypothèses emboitées. Ces tests vont permettre de faire du choix de modèle sur la fonction potentiel.

Les données sur le maïs ont été analysées en utilisant le cadre statistique décrit avant. La conjontion du modèle de Gibbs et le cadre fonctionnel d'estimation du potentiel se prête bien à une description assez fine des données. Nous avons mis en évidence qu'une démarche descendante pour le choix de modèles n'aboutissait pas car l'existence de trop de composantes pose des problèmes. Par contre l'approche ascendante permet de faire une description acceptable des données et de raisonner le choix de modèle. Nous avons mis en évidence une procédure alterna-

tive d'estimation et de choix basée sur une analyse spectrale et du maximum de vraisemblance.

Références

- Billiot J.M. & Goulard M. 2001. An estimation method of the pair potential function for Gibbs point processes on spheres. *Scandinavian Journal of Statistics*, 28, 185-203.
- Billiot J.M. (1995). Estimation dans les modèles spatiaux de Gibbs : synthèse bibliographique. *Publications de l'I.S.U.P. 39*, fasc. 2, 3-33.
- Carter D.S., Prenter P.M. (1972). Exponential spaces and counting processes. *Zeitschrift für Wahrsheinlichkeitstheorie und verwandte Gebiete*, 21, 1-19.
- Chadœuf J., Goulard M., Pellerin S. (1993). A Gibbs point process on finite series of circles : the insertion of the primary roots of maize around the stem. *J. Appl. Stat.* 20, no 1, 177-185.
- Daley D.J. Vere-Jones D. (1988). An Introduction to the theory of Point Processes. Springer-Verlag, New York, 1988.
- Diggle P.J., Gates D.J., Stibbard A. (1987). A non parametric estimator for pairwise-interaction point process. *Biometrika* 74, 763-770.
- Fiksel T. (1988). Estimation of interaction potentials of Gibbsian point processes. *Statistics* 19, 1, 77-86.
- Geyer C.J. (1999). Likelihood inference for spatial point processes. In O.E. Barndorff-Nielsen, W.S. Kendall and M.N.M. van Lieshout, editors, Stochastic Geometry : Likelihood and Computation. Chapman and Hall, London, 1999. (To appear).
- Geyer C.J., Moller J. (1994). Simulation procedures and likelihood inference for spatial point processes. *Scand. J. Stat.* 21, 359-373.
- Glötzl E., Rauschenschwandtner B. (1981). On statistics of Gibbsian point processes. In : *The First Pannonian Symposium on Mathematical Statistics.* Lectures Notes in Statistics, 8. Springer.
- Heikkinen J., Penttinen A. (1994). Bayesian Smoothing in the Estimation of the Pair Potential Function for Gibbsian Point Processes. *Preprints from the Department of Statistics* 17, University of Jyväskylä.
- Jensen J.L., Moller J. (1991). Pseudolikelihood for exponential family of spatial processes. *Ann. App. Prob.* 1, 445-461.
- Ogata Y., Tanemura M. (1981). Estimation of interaction potentials of spatial point patterns through the maximum likelihood procedure. *Ann. Inst. Stat. Math. B* 33, 315-338.
- Pellerin S., Trendel R., Duparque A. (1990). Relation entre quelques ca-

ractères morphologiques et la sensibilité à la verse en végétation de maïs. (Zea mays L.). *Agronomie* 10, 439-446.
– Pellerin S., Tricot F. Chadœuf J. (1989). Disposition des racines adventives autour de la tige de maïs (Zea mays L.). *Agronomie* 9, 859-866.
– Penttinen A. (1984). Modelling Interactions in spatial point patterns : Parameter Estimation by the Maximum Likelihood Method. *Jyväskylä Studies in Computer Science, Economics and Statistics* 7, 107p.
– Ripley B.D. (1977). Modelling Spatial Patterns (with discussion), *J. Roy. Statist. Soc. Ser. B* 39, 172-212.
– Stoyan D., Kendall W.S., Mecke J. (1995). *Stochastic Geometry and its Applications.* Wiley Series in Probability and Mathematical Statistics, 2nd Edition. Akademie-Verlag, Chichester, Berlin, 436p.

50

.

Vitesse optimale du cas i.i.d. pour l'estimation non-paramétrique de la densité invariante d'un système dynamique chaotique

Salim LARDJANE
Université de Bretagne Sud & CREST-ENSAI
CREST-ENSAI, Laboratoire de Statistique et Modélisation
Rue Blaise Pascal, 35170 Bruz
e-mail : lardjane@ensai.fr

Résumé

Les systèmes dynamiques déterministes sont souvent utilisés pour modéliser des phénomènes évolutifs en écologie, en physique et en économie [6,7,8].

Je m'intéresse ici à l'estimation non-paramétrique de la densité invariante d'un système dynamique appartenant à une classe de transformations chaotiques. Cette densité décrit le comportement à long terme du système ou encore un état d'équilibre statistique de celui-ci [1,11].
Les systèmes considérés sont unidimensionnels et évoluent en temps discret ; leurs orbites $(x_t)_{t\in\mathbb{N}}$ sont définies par des équations de la forme $x_{t+1} = S(x_t)$ où x_0 est l'état initial du système et où S est sa loi d'évolution, également appelée *transformation itérée* [5].

Je commence par démontrer diverses inégalités de covariance [4]. Je démontre ensuite que le problème est *équivalent* à celui de l'estimation de la densité marginale d'un processus stochastique stationnaire $(X_t)_{t\in\mathbb{N}}$ tel que $X_{t+1} = S(X_t)$ où X_0 est une variable aléatoire fixée [4]. De tels processus sont appelés *processus dynamiques* et sont hautement dépendants. Il ne peuvent, en particulier, être supposés mélangeants [2,3].

Dans un premier temps, j'utilise cette équivalence pour étudier les propriétés de l'estimateur de Parzen-Rosenblatt de la densité marginale d'un processus dynamique $(X_t)_{t\in\mathbb{N}}$ basé sur l'observation de $X_0, X_1, \ldots, X_{n-1}$. Celui-ci est défini, en tout point x, par

$$f_n(x) = \frac{1}{nh_n}\sum_{t=0}^{n-1} K(\frac{x - X_t}{h_n}),$$

où K est le noyau uniforme $\chi_{[-1/2,1/2]}$ et où $(h_n)_{n\in\mathbb{N}}$ est une suite de fenêtres décroissant vers 0 [9,10].

Je démontre que l'estimateur peut, *dans certains cas*, converger en moyenne quadratique *avec la vitesse optimale du cas i.i.d.*, pour les même suites de fenêtres que dans le cas i.i.d [4].

Dans un deuxième temps, j'applique ce résultat aux systèmes dynamiques chaotiques considérés, et donne une interprétation pratique de l'erreur quadratique moyenne dans ce contexte [4].

Références

1 G. D. Birkhoff, G. D. (1931). Proof of the ergodic theorem. . *Proc. Nat. Acad. Sci. USA*, **17**, 656-660.

2 Bosq, D. (1995). Optimal asymptotic quadratic error of density estimators for strong mixing or chaotic data. *Statistics and Probability Letters*, **22**, 339-347.

3 Bosq, D. (1998). *Nonparametric statistics for stochastic processes : estimation and Prediction*. Lecture Notes in Statistics, **110**. Springer.

4 Lardjane, S. (2002).Optimal speed nonparametric density estimation for one-dimensional dynamical systems. *Série des document de travail du CREST*, INSEE, **16**. Disponible sur `www.crest.fr`.

5 Lasota, A. and Mackey, M. (1994). Chaos, *Fractals and Noise : Stochastic Aspects of Dynamics*. Springer-Verlag.

6 May, R. (1976). Simple mathematical models with very complicated dynamics. *Nature*, **261**, 459-467.

7 Medio, A. (1992). *Chaotic Dynamics : Theory and applications to Economics*. Cambridge.

8 Neimark, Y. and Landa, P. (1992). *Stochastic and Chaotic Oscillations*. Kluwer.

9 Parzen, E. (1962). On estimation of a probability density function and mode. *Ann. Math. Stat.*, **33**, 1065-1076.

10 Rosenblatt, M. (1956). Remarks on some non parametric estimates of a density function. *Ann. Math. Stat.*, **27**, 832-837.

11 Ruelle, D. (1989). *Chaotic evolution and strange attractors*. Cambridge.

Estimation de la fonction du taux d'occurence d'événements ponctuels

Christophe BONALDI et Nicolas MOLINARI *
* Adresse pour correspondance :
Laboratoire de Biostatistique
Institut Universitaire de Recherche Clinique
641 avenue Gaston Giraud, 34093 Montpellier.
e-mail : molinari@iurc.montp.inserm.fr

Résumé

Notons $X_1, \cdots, X_N$ des variables aléatoires indépendantes et identiquement distribuées représentant les dates d'occurrence de N événements sur un intervalle de temps $(0, T)$. Nous souhaitons tester l'hypothèse nulle de répartition uniforme des événements contre l'hypothèse alternative correspondant à la présence d'agrégats ou de clusters sur des sous-intervalles de $(0, T)$. Ce type d'étude est particulièrement intéressant en épidémiologie pour l'étude de maladies rares comme par exemple la leucémie [1].
De nombreux tests ont été développés afin de déterminer la présence d'un unique cluster [2] [3]. En revanche, il n'existe que peu de tests permettant de déterminer plusieurs clusters [4] et ils ne sont souvent valables que pour une population à risque constante tout au long de l'étude [5]. Cette dernière hypothèse n'est que peu réaliste. En effet, que ce soit dû à des phénomènes saisonniers ou naturels, la taille d'une population à risque change au cours d'études qui durent plusieurs années.
Dans une première approche [6], nous proposons de transformer les observations de manière à se retrouver dans un contexte de régression. Notons $X_{(1)}, \cdots, X_{(N)}$ les statistiques d'ordre associées à $X_1, \cdots, X_N$. Intéressons-nous aux données $(i, Y_i)_{i=1,\cdots,N}$ où $Y_i = X_{(i)} - X_{(i-1)}$ représente la distance entre deux événements consécutifs. Sous l'hypothèse de répartition uniforme, on peut régresser ces données par une constante, la moyenne des Y_i. Supposons maintenant que $X_{(k)}, \cdots, X_{(k+l)}$ soient regroupées en cluster, les distances $Y_{k+1}, \cdots, Y_{k+l}$ seront en moyenne plus petites que les autres Y_i. Ainsi, un modèle approprié serait une régression par morceaux. Le calcul des bornes de chaque clusters se fait en résolvant un problème des moindres carrés. Pour prendre en compte l'évolution de la population à risque, les données sont modifiées en $(i, \breve{y}_i) = (i, y_i \times R(x_i))$ pour $i = 1, \cdots, N$. La fonction

$R(t)$ donne le taux d'évolution de la population au cours du temps t.
Cette première approche peut être généralisée par la modélisation de la fonction de taux d'un processus de Poisson. L'introduction de covariables est alors possible et l'estimation du taux et des cœfficients de régression se fait par maximum de vraisemblance [7]. Des critères classique de choix de modèle permettent alors de déterminer la présence d'un agrégat.
Une autre méthode d'estimation du taux d'occurence est d'utiliser des méthodes MCMC. Les taux sont modélisés par des fonctions constantes par morceaux et estimés en deux temps. Une première étape se fait par MCMC à saut réversible [8] afin d'estimer le taux a posteriori. Dans une seconde étape, on calcule les estimateurs de Bayes conditionnellement au nombre de sauts en utilisant un algorithme de recuit-simulé. Il s'agit enfin de déterminer le nombre de sauts et de comparer le taux estimé à celui correspondant à une répartition uniforme de dates d'occurence.
L'ensemble de ces différentes approches a été utilisé sur des données de cancers de la thyroïde dans plusieurs département français et des cas d'hémoptysie dans la région niçoise.

Références

1 Ederer, F., Myers, E. & Mantel, N. (1964), A statistical problem in space and time : do leukemia cases come in clusters ?, *Biometrics*, 20, 623-626.
2 Naus, J. I. (1965), The distribution of the size of the maximum cluster of points on a line, *JASA*, 60, 532-538.
3 Tango, T. (1984), The detection of disease clustering in time, *Biometrics*, 40, 15-26.
4 Glaz, J. & Naus,J. (1983), Multiple clusters on the line, *Comm. Statist.-Theor. Meth.*, 12, 1961-1986.
5 Weinstock, M. A. (1981), A generalized scan statistic test for the detection of clusters, *Int. J. Epidemiol.*, 10, 289-293.
6 Molinari, N., Bonaldi, C. & Daurès, J.P. (2001), Multiple temporal cluster detection, *Biometrics*, 57, 577-583.
7 Cox, D. R. & Lewis, P. A. W. (1966), *The Statistical Analysis of Series of Events*, Methuen, London.
8 Green, J. (1995), Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, 82, 711-732.

Estimation fonctionnelle des modes conditionnels. Cas d'une chaine de Markov incluse à densite conditionnelle bimodale

Abbes RABHI* et Abderrahmane YOUSFATE
* Adresse pour correspondance :
Laboratoire des Mathématiques, Département des Mathématiques
Faculté des Sciences
UDL, Sidi bel Abbès 22000, Algérie.
e-mail : yousfate_a@yahoo.com

Résumé

Dans ce travail, on s'intéresse à l'estimation des modes d'une densité conditionnelle d'une chaîne de Markov synthétique incluse issue d'une file d'attente $GI/GI/1$. Pour ce faire, on donne d'abord un estimateur fonctionnel de la densité de l'opérateur de transition en améliorant certains résultats dus à [4]. Ensuite, sous des conditions peu restrictives, on étudie les propriétés de l'estimateur des modes conditionnels. Enfin on donne des expressions générales des lois qui interviennent dans l'étude pour faciliter la simulation.

1. Introduction

Dans ce travail, on considère un système (A, B) (A ou B pouvant être vide(s)) recevant des éléments en A et les expulsant par B tel que A induise à tout instant t un ensemble $A(t)$ ayant une structure d'ordre issue de l'ordre d'arrivée des éléments dans A (les arrivées sont indépendantes entre elles.) Les inter arrivées successives dans A suivent une loi générale. Le système A se vide (selon l'ordre des arrivées) dans le système B qui contient au plus un (ayant transité par A.) La durée de séjour d'un élément dans B suit une autre loi générale. On considère également que les durées de séjour sont indépendantes entre elles et que la loi des inter-arrivées et la loi de séjour dans B sont indépendantes. En considérant que le système (A, B) soit vide à l'instant t_0, on montre que les durées d'attente et les durées d'innoccupation du système (A, B) sont des chaînes de Markov (incluses), les instants à considérer sont les instants d'arrivée en A et les instants de sortie de B qui sont des temps d'arrêt. On considère qu'il n'y ait pas de simultanéité

entre les instants d'arrivée et les instants de sortie du système (A, B) ainsi décrit.

On note A_i (respectivement B_i) le temps d'inter arrivées dans A entre le $(i-1)$-ème et le i-ème élément (respectivement le temps de séjour de l'élément i dans B). La loi de A_i est notée F_A et la loi de B_i est notée F_B.

Le temps d'attente du $(i+1)$-ème élément est obtenu par l'équation de récurrence qui suit :

$$\mathcal{W}_{i+1} = (\mathcal{W}_i + B_i - A_{i+1})_+$$

si $\mathcal{W}_{i+1} = 0$, $(\mathcal{W}_i + B_i - A_{i+1})_-$ mesure le temps d'innoccupation du système que l'on note I_{i+1}.

$$\begin{aligned} I_{i+1} &= (\mathcal{W}_i + B_i - A_{i+1})_- \\ &= \left((\mathcal{W}_{i-1} + B_{i-1} - A_i)_+ + B_i - A_{i+1}\right)_- \\ &= (\mathcal{W}_{i-1} + B_{i-1} - A_i + I_i + B_i - A_{i+1})_- \end{aligned}$$

En considérant $(\mathcal{W}_i, I_i)_{i\in\mathbb{N}}$, on montre aisément que c'est une chaîne de Markov ayant la particularité si $\mathcal{W}_i$ est positif, I_i est nul et vice versa. Cela nous permet de construire une variable scalaire sur $\mathbb{R}$, $Z_i = \mathcal{W}_i - I_i$ équivalente à $(\mathcal{W}_i, I_i)$. Cette variable a pour indice des temps d'arrêt et vérifie les propriétés d'une chaîne de Markov.

On suppose que $(Z_i)_i$ est ergodique (c'est vérifié dés que $\mathbb{E}(B_i) < \mathbb{E}(A_i)$, voir [5]) et que sa densité ψ est continue, bornée et à support dans $\mathbb{R}$. La mesure stationnaire sera notée ν et la densité conditionnelle sera notée $p(\omega, \omega')$.

Pour utiliser un estimateur fonctionnel de $p(\omega, \omega')$, on utilise un noyau de Nadaraya-Watson K positif, borné, intégrable, à support compact noté $[\rho_1, \rho_2]$ tel que $\rho_1\rho_2 < 0$ et minoré par $\varepsilon > 0$ sur le support. En notant $K\left(\dfrac{x}{h_n}\right) = K_{h_n}(x)$ l'estimateur s'écrit :

$$p_n(\omega, \omega') = \frac{\sum_{i=1}^n K_{h_n}(\omega - Z_i)\, K_{h_n}\left(\omega' - Z_{i+1}\right)}{\sum_{i=1}^n K_{h_n}(\omega - Z_i)}\,.$$

L'estimateur p_n a été déjà utilisé par [1] et [4]. Notons aussi que [7] a utilisé cet estimateur pour étudier la densité conditionnelle en utilisant un n-échantillon i.i.d. Si on est intéréssé juste par la prévision, on utilise l'estimateur suivant :

$$p_n(\omega, \omega') = \frac{\sum_{i=1}^n K_{h_n}(\omega - Z_i)\, Z_{i+1}}{\sum_{i=1}^n K_{h_n}(\omega - Z_i)}\,.$$

qui a été utilisé par [2]. Considérons les hypothèses :

(H.1) $\mathbb{E}(B) < \mathbb{E}(A)$ (*hypothèse d'ergodicité*) *[5].*

(H.2) $\forall x \in \mathbb{R}$, $\forall \Gamma(x)$ *mesurable ; tel que* $0 < \nu(\Gamma(x)) < 1$, *la fermeture de* $\Gamma(x)$ *est connexe et contient* x. *De plus pour un certain* $\alpha \in]1/2, 1]$ *et un* $\beta > 0$,

$\exists\ m \in \mathbb{N}^*;\ m \leq \dfrac{1}{\alpha(1-\nu(\Gamma(x)c))}$
tel que $\forall z \in \Gamma(x),\ \displaystyle\int_{\Gamma(x)} p(z,m,y)\,dy \leq \beta\,\nu(\Gamma(x))$.
Cette hypothèse veut dire que si Z_i prend une valeur dans un voisinage de x, notée $\Gamma(x)$ mesurable à fermeture connexe tel que $0 < \nu(\Gamma(x)) < 1$, alors il doit quitter la fermeture de $\Gamma(x)$ avant $\dfrac{1}{\alpha\nu(\Gamma(x)^c)}$ avec une probabilité supérieure à $1 - \beta\,\nu(\Gamma(x))$. Sans perdre de généralité, nous pouvons considérer par la suite que $\Gamma(x)$ est fermé et connexe.

(H.3) Le noyau K est borné, intégrable, à support compact noté $[\rho_1, \rho_2]$ tel que $\rho_1\rho_2 < 0$ et minoré par $\varepsilon > 0$ sur le support.

(H.4) $\psi(x)$ est continue, strictement positive, bornée et de classe $\mathcal{C}^2$. De plus $p(x,y)$ est bornée de classe $\mathcal{C}^2$ vérifiant $\left|\dfrac{\partial p(x,y)}{\partial x}\right|$ et $\left|\dfrac{\partial p(x,y)}{\partial y}\right|$ est bornée relativement à x et y.

(H.5) Pour chaque x, on note $\Gamma_n(x) =]x - \rho_2 h_n, x - \rho_1 h_n[$. Quand $n \to \infty$, $h_n \to 0$ de telle sorte que $n\,(\nu(\Gamma_n(x)))^2\,h_n^2 \to \infty$.

2. Les modes de convergence

→ La convergence uniforme p.s.
Sous les hypothèses (H.1)-(H.4) et si $h_n \longrightarrow 0$ $nh_n \longrightarrow \infty$, alors on a : $\forall \omega \in \mathbb{R}$

$$\sup_{\omega' \in \mathbb{R}} \left| p_n(\omega,\omega') - p(\omega,\omega') \right| \longrightarrow 0, \qquad \nu\text{-p.s.}$$

Ce résultat généralise celui obtenu par [4].

→ La convergence uniforme presque complète.
Sous les hypothèses (H.1)-(H.5) et si K vérifie la condition de Hölder d'ordre b

$$\left(\exists b > 0,\ \exists C < \infty,\ \forall \omega,\ \omega' \in \mathbb{R} :\ \left|K(\omega) - K(\omega')\right| \leq C\left|\omega - \omega'\right|^b\right),$$

alors on a :

$$\sup_{(\omega,\omega') \in \mathbb{R}^2} |p_n(\omega,\omega') - p(\omega,\omega')| \longrightarrow 0,\ \text{p.co.}$$

En s'ispirant de la décomposition faite par [3], on utilise les fonction aléatoires :

$$\psi_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}\left(\omega - \mathcal{W}_i\right) \quad \text{et}$$

$$\varphi_n(\omega,\omega^{'}) = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}\left(\omega - \mathcal{W}_i\right) K_{h_n}\left(\omega^{'} - \mathcal{W}_{i+1}\right)$$

estimant respectivement la densité ψ de la loi stationnaire et la densité de probabilité conjointe, ainsi $p_n(\omega,\omega^{'})$ peut être représenté comme rapport de $\dfrac{\varphi_n(\omega,\omega^{'})}{\psi_n(\omega)}$, estimant $p(\omega,\omega^{'})$ qui n'est autre que le rapport de $\varphi(\omega,\omega^{'})$ et $\psi(\omega)$. Pour vérifier l'assertion du théorème , on utilise les résultats suivants.

$$\sup_{\omega^{'}\in\mathbb{R}} \left|\mathbb{E}\varphi_n(\omega,\omega^{'}) - \varphi(\omega,\omega^{'})\right| \longrightarrow 0,$$
$$\sup_{\omega\in\mathbb{R}} \left|\mathbb{E}\psi_n(\omega) - \psi(\omega)\right| \longrightarrow 0.$$

Sous le hypothèses du théorème précédent on a :

$$\sup_{\omega^{'}\in\mathbb{R}} \left|\mathbb{E}\varphi_n(\omega,\omega^{'}) - \varphi_n(\omega,\omega^{'})\right| \longrightarrow 0, \text{ p.co.}$$
$$\sup_{\omega\in\mathbb{R}} \left|\mathbb{E}\psi_n(\omega) - \psi_n(\omega)\right| \longrightarrow 0, \text{ p.co.}$$

Théorème d'Azuma 1967. Si $(Z_i)_{i\in\mathbb{N}}$ est une différence martingale telle qu'il existe un réel positif δ qui vérifie $|Z_i| \leq \delta$ p.s. Alors : $\forall \varepsilon > 0$

$P\left[\left|\sum_{i=1}^{n} Z_i\right| > \varepsilon\right] \leq 2\exp\left(-\dfrac{n\varepsilon^2}{2\delta^2}\right)$

3. Les vitesses de convergence

Dans cette partie en premier lieu on va donner les vitesses de convergence au sens de la norme L^∞ et la convergence uniforme presque complète pour l'estimateur de la densité de la transition.

Théorème. Si le processus de Markov est à valeur dans $\mathbb{R}$ et si les hypothèses $(H.1)$-$(H.5)$ sont vérifiées, alors

$$\sup_{(\omega,\omega^{'})\in\mathbb{R}^2} \left|p_n(\omega,\omega^{'}) - p(\omega,\omega^{'})\right| = O\left(h_n^2\right).$$

Sous les mêmes hypothèses on a :

$$\sup_{\omega^{'}\in\mathbb{R}} \left|p_n(\omega,\omega^{'}) - p(\omega,\omega^{'})\right| = O\left(h_n^2\right) + O\left(\frac{\log n}{nh_n}\right), \text{ p.co.}$$

4. Estimation des modes conditionnels

Supposons que $p_n(\omega,\omega^{'})$ admette un maximum unique dans $\mathbb{R}_+$ en un point θ_+ et un autre maximum unique dans $\mathbb{R}_-$ noté θ_-. On suppose que θ_+, θ_- et $p(\omega,\omega^{'})$ sont inconnus et que l'on dispose d'un estimateur $p_n(\omega,\omega^{'})$ de $p(\omega,\omega^{'})$ basé sur n observations. Si $p_n(\omega,\omega^{'})$ possède lui même deux maxima en θ_{+n} et θ_{-n}, il est naturel d'estimer θ_+ (respectivement θ_-) par θ_{+n} (respectivement θ_{-n}).

Le but de cette partie est de montrer la convergence en probabilité ainsi que la convergence presque complète d'un estimateur vectoriel des modes conditionnels construits à partir d'un estimateur convenable de la densité de la transition et de donner une vitesse de convergence de $(\theta-\theta_n)$ où $\theta=(\theta_+,\theta_-)$ et $\theta_n=(\theta_{+n},\theta_{-n})$.

Avec les notations précédentes, on a les résultats suivants qui établissent les modes de convergence de l'estimateur de θ. Sous les hypothèses (H.1)-(H.5) et si $nh_n^2 \underset{n\longrightarrow\infty}{\longrightarrow} \infty$ et $p(\omega,\omega^{'})$ est uniformement continue, alors quand $n\longrightarrow\infty$, on a, $\forall\varepsilon>0$:

$$P\left(\sup_{\omega^{'}}\left|p_n(\omega,\omega^{'})-p(\omega,\omega^{'})\right|<\varepsilon\right),$$

$$P\left(|\theta_n-\theta|<\varepsilon\right)\longrightarrow 1.$$

Si $nh_n^2\nu(\Gamma_n(x))^2 \underset{n\longrightarrow\infty}{\longrightarrow} \infty$ et $\forall\omega,\ \omega^{'}\in\mathbb{R}$, $p(\omega,\omega^{'})$ admet un couple de modes unique et qu'elle est uniformement continue. Si de plus

$$\sup_{\omega^{'}}\left|p_n(\omega,\omega^{'})-p(\omega,\omega^{'})\right|\longrightarrow 0,\quad \text{p.co.},$$

alors

$$|\theta_n-\theta|\longrightarrow 0,\quad \text{p.co.}$$

Pour la démonstration (facile), on peut s'inspirer de celle de [6].

5. Simulations

Pour faire des simulations stochastiques on utilise la relation (vérifiable aisément)

$$P\left(Z_{n+1} < z \ / Z_n = u\right) = \begin{cases} 1 - F_A * \overline{F}_B(u - z) & \text{si } \ u >_0 \\ \\ 1 - F_A * \overline{F}_B(-z) & \text{si } u \leq 0 \end{cases}$$

où F_A est la fonction de répartition de A et $\overline{F}_B$ est la fonction de répartition du symétrique de B.

Pour le choix du noyau, on utilise un noyau de Nadaraya-Watson asymétrique approprié ou une suite de noyaux appropriés.

Références

1 G. Roussas, (1969). Nonparametric estimation of the transition distribution function of a Markov process. *Annals of Math. Stat.*, **40**, pp 1386-1400.

2 G. Collomb G, (1984). Propriétés de convergence presque complète du prédicteur à noyau. *Zeitschrift für Wahrscheinlichkeisttheorie und verwandte Gebiete*, **66**, 441-440.

3 F.Ferraty F. et P. Vieu, (2001). *2001 "Statistique fonctionnelle : modèles de régression pour variables uni, multi et infiniment dimensionnées"*. Publications du LSP 2001-03, Toulouse, France.

4 A. Laksaci & A. Yousfate, (2002). Estimation fonctionnelle de la densité de l'opérateur de tranition d'un processus de Markov à temps discret. *C. R. Acad. Sci. Paris, Ser. I*, **334**, 1179-1202.

5 J. Neveu, (1983). *Construction des files d'attente stationnaires*. Notes on control and information Sciences **60**. Springer-Verlag, 31-41.

6 P. Vieu, (1986). A note on density mode estimation. *Statistics & Probability Letters*, **26**, pp 297-307.

7 E. Youndjé, (1993). Propriété de convergence de l'estimateur à noyau de la densité conditionnelle. *Rev. Roumaine Math. Pure Appl.*, **41**,(7-8), 535-566

Sur la convergence forte dans l'estimation de la densité spectrale pour les processus stationnaires à temps continu après échantillonnage du temps.

Mustapha RACHDI
Université Pierre Mendès-France,
IMAG-LMC UMR 5523 et LabSAD, Dpt. IMSS-BSHM
BP 47, 38040 Grenoble Cedex 09
e-mail : Mustapha.Rachdi@imag.fr

Résumé

L'analyse de Fourier des données et l'estimation de la densité spectrale pour les processus à temps continu ont prouvé leur utilité en communication [6] (dans le filtrage linéaire et la théorie de prédiction), en sismologie [4] (pour déterminer la nature d'un événement sismique), en océanographie [9] (pour l'étude des ondes océaniques), et dans divers domaines des sciences physiques, médicales, ...

Dans la statistique des processus à temps continu, les données sont souvent collectées en utilisant un schéma d'échantillonnage. Différents échantillonnages du temps peuvent être employés. Dans cette présentation, nous considérons deux types d'échantillonnages : l'échantillonnage périodique et l'échantillonnage aléatoire. Ce choix est motivé par les raisons suivantes : dans la théorie classique de l'estimation spectrale, le périodogramme se calcule à partir des observations du processus X sur $[0, T]$ où $T > 0$, par :

$$I_{_T}(\lambda) = 1/(2\pi T)| \int_0^T X(t)\ \exp(-i\,\lambda\,t)\,dt|^2,\ \forall \lambda \in \mathbb{R}.$$

Cet estimateur n'est pas convergent, mais en le lissant, on construit un estimateur de type noyau $\widehat{\phi}_{_X}$ qui est asymptotiquement consistant sous certaines conditions. La convergence en moyenne quadratique de $\widehat{\phi}_{_X}$ a été obtenue dans [2] et [11], et avec échantilllonnage du temps dans [7] et [8]. La convergence presque complète (p.co.) *sans* échantillonnage du temps a été obtenue dans [3].

Dans les applications pratiques, les observations de X ne sont pas obtenues sous une forme analytique, alors l'intégrale

$$\int_0^T X(t)\ \exp(-i\omega t)\ dt,$$

ne peut être calculée numériquement. Ceci constitue un problème majeur pour calculer le périodogramme. A cette fin, X est observé à des instants $\{t_n\}_{n\in\mathbb{Z}}$ et les observations sont

$$\{X(t_n), n = 1, \ldots, N\},\ N \in \mathbb{N}.$$

Ces instants d'échantillonnage $\{t_n\}_{n\in\mathbb{Z}}$ sont à déterminer. Dans cet exposé, nous choisissons dans un premier temps, l'échantillonnage périodique. Ce choix introduit le phénomène de repliement des ondes (aliasing). Dans ce cas, la densité spectrale ϕ_X se calcule à partir de la densité spectrale $\phi_{\widetilde{X}}$ correspondante au processus échantillonné

$$\widetilde{X} = \{X(n/\epsilon)\}_{n\in\mathbb{Z}},\ \epsilon > 0,$$

si le processus stochastique X est bande-limité (ϕ_X est à support dans $[-\pi\epsilon, \pi\epsilon]$) [8]. Nous construisons ensuite l'estimateur de ϕ_X, et établissons sa convergence p.co. et sa vitesse de convergence. La condition sur le support de ϕ_X semble être restrictive. C'est pour cela que dans un deuxième temps, nous adoptons l'échantillonnage aléatoire qui pallie le phénomène d'aliasing (voir [2] et [15]). Nous construisons ensuite l'estimateur de la densité spectrale, puis nous établissons sa convergence uniforme p.co. et donnons sa vitesse de convergence. Notons que la convergence en moyenne quadratique de cet estimateur a été établie dans [7].

Par ailleurs, le lissage du périodogramme introduit un paramètre dit largeur de fenêtre spectrale. Ce paramètre joue un rôle crucial dans la vitesse de convergence de $\widehat{\phi}_X$. Nous discuterons donc le choix optimal de ce paramètre (voir [12]).

Références

1 Bosq, D. (1998). *Nonparametric statistics for stochastic processes : Estimation and Prediction, Second Edition.* Lecture notes in statistics, **110**, Springer-Verlag.

2 Brillinger, D.R. (1975). *Time series analysis and theory.* Holt, Rienhart and Winston, New York.

3 Carbon, M. (1981). Sur la convergence uniforme presque sûre des estimateurs de la densité spectrale des processus stationnaires et mélangeants. application à l'erreur de prédiction linéaire. *C. R. Acad. Sc. Paris*, **292**, 95-98.

4 Carpenter, E.W. (1967). Explosions seismology. *Science*, **147**, 363-373.

5 Charlot, F. and Rachdi, M. (2003). On the statistical properties of a stationary process sampled by a stationary point process. *Stat. & Probab. Lett.*,

to appear.

6 Davenport, W.B. and Root, W.L. (1958). *An introduction to the theory of random signals and noise*. McGraw-Hill, New York.

7 Lii, K.S. and Masry, E. (1994). Spectral estimation of continuous-time stationary processes from random sampling. *Stoch. proc. and their appli.*, **52**, 39-64.

8 Masry, E., Klamer, D. and Mirabile, C. (1978). Spectral estimation of continuous-time processes : performance comparison between periodic and Poisson sampling schemes. *IEEE Trans. on Auto. Control*, **AC-23**(4), 679-685.

9 Moore, M.I., Thomson, P.J. and Shirtcliffe, T.G. (1988). Spectral analysis of ocean profiles from unequally spaced data. *J. Geophysical Res*, **93**, 655 -664.

10 Politis, D. and Romano, J.P. (1999). *Resampling*. Springer, New York.

11 Priestley, M.B. (1981). *Spectral analysis and time series*, vol 1, (Univariate series). Academic Press.

12 Rachdi, M. (1998). Cross-validated choice of the spectral bandwidth for a continuous stationary process. *C. R. Acad. Sci., Sér. 1, Math.*, **327**(8), 777-780.

13 Rachdi, M. and Sabre, R. (1999). The optimal choice of the spectral bandwidth for a random field. *Trait. Signal.*, **15**(6), 569-575.

14 Rachdi, M. and Monsan, V. (1999). Asymptotic properties of p-adic spectral estimates of second order. *Jr. Comb. Info. & Sys. Sci.,*, **24**(2), 113-142.

15 Shapiro, H.S. and Silverman, A.R. (1960). Alias-free sampling of random noise. *Soc. Indust. Appl. Math*, **8**(2), 225-48.

16 Yoshihara, K. (1992). *Weakly dependent stochastic sequences and their applications*, volume 1 : Summation theory for weakly dependent sequences. Sanseido, Tokyo.

64

.

Propriétés extrémales des valeurs singulières d'un opérateur compact et application en analyse factorielle

Jean Jacques TÉCHENÉ
Université de Pau et des Pays de l'Adour
e-mail : jean-jacques.techene@univ-pau.fr

Résumé

Les analyses factorielles sont inspirées de manière plus ou moins directe, de techniques relativement anciennes (Hotteling, 1933) élaborées dans un cadre probabiliste en dimension finie en vue d'applications à des problèmes de Statistique Inférentielle puis adaptées récemment à des problèmes de Statistique Multivariée. Ces analyses ont été, à l'origine, définies comme des méthodes itératives (ou "pas à pas"), et alors soumises à des conditions relativement exigeantes d'existence. Mais un cadre probabilitse plus général et une approche globale s'imposent si l'on veut mieux comprendre les propriétés asymptotiques de ces analyses et leur stabilité par échantillonnage ou discrétisation.

Nous cherchons à obtenir, à partir des caractères extrémaux des valeurs singulières d'un opérateur compact, des critères globaux (i.e. non itératifs) d'analyses factorielles linéaires d'une probabilité définie sur un espace de Hilbert réel séparable ou d'une fonction aléatoire réelle. L'idée de chercher des critères globaux d'analyses factorielles linéaires équivalents à l'ACP d'une famille finie de v.a.r. revient à Rao (1964), Darroch (1965) puis Okamoto (1969). Notre étude constitue une extension de cette recherche dans un cadre probabiliste sensiblement plus général.

Notre approche se fonde sur la recherche, dans le cadre hilbertien, de critères globaux de réduction canonique d'un opérateur compact non nécessairement auto-adjoint, desquels nous déduisons des caractères d'extrémalité des valeurs singulières d'un tel opérateur. Ces caratères généralisent, pour certains, des propriétés connues dans un cadre matriciel mais avec une formulation sensiblement plus précise, et pour d'autres, des résultats établis par Göhberg and Krejn (1969).

Ces critères reposent sur deux problèmes importants d'optimisation définis à l'aide de la famille des normes symétriques $||.||_p$, $p = 1, \ldots p = +\infty$ sur l'espacde

des opérateurs compacts sur un espace de Hilbert complexe séparable, et que nous résolvons à l'aide d'une extension dans ce cadre du Théorème de Poincaré de séparation des valeurs d'une matrice hermitienne (1890) ou de résultats établis par rao (1979).

Etant donnés deux espaces de de Hilbert complexes séparables H et H' et un opérateur compact T de H dans H', le premier problème consiste en la recherche d'un sous-espace F de dimension finie q fixée de H tel que $||T||_F$ ou $||PoT_F||_p$ soit maximum, P et T_F désignant respectivement le projecteur ortoghonal sur F et la restriction de T à F, ou plus généralement en la recherche d'un couple (F, F') de sous-espaces de dimensions finies $q' \geq q$ tel que $||P'oToP||_p$ soit maximum. La solution est indépendante du réel $p \geq 1$ choisi : il faut et il suffit que F et F' soient engendrés respectivement par q premiers vecteurs propres de T^*oT et q' premiers vecteurs propres de ToT^*.

Le second problème consiste en la recherche d'un opérateur U de rang fini fixé tel que $||T - U||_p$ soit minimum, généralisant le cas $p = 2$ bien connu dans le cadre euclidien et matriciel (Rao, 1964). Nous établissons notamment que, pour chaque $p \geq 1$, ce minimum est atteint si et seulement si U est le développement de Schmidt d'ordre q de T, complétant ainsi sensiblement un théorème établi par Göhberg and Krejn (1969). Nous accordons également une attention particulière au cas $p = \infty$ complétant encore sur ce point le travail de Göhberg and Krejn (1969).

Nous en déduisons plusieurs carctères extrémaux des valeurs propres d'un opérateur compact auto-adjoint positif ou des valeurs singulières d'un opérateur compact non auto-adjoint, le premier de ces résultats ayant été établi par Wielandt (1955) dans le cadre euclidien. De tous ces caractères d'extrémalité, nous déduisons des critères non itératifs d'analyses factorielles linéaires d'une probabilité sur un espace de Hilbert réel séparable ou d'une fonction aléatoire réelle non nécessairement réduite à une famille finie de v.a.r. Outre l'avantage d'assurer dans tous les cas l'existence de l'analyse et de fournir un moyen global de l'obtenir, une telle approche permet de dégager plusieurs familles de critères, plus forts, équivalents ou plus faibles que l'ACP définie dans un cadre probabiliste général par Dauxois et Pousse (1976).

Références

1 J.N. Darroch, (1965). An extremal property of principal components. *Ann.*

Math. Statist., **36**, 1579-1582.

2 J. Dauxois et A. Pousse, (1976). *Les analyses factorielles en calcul des probabilités et en statistique*, Thèse d'état, Univ. P. Sabatier, Toulouse.

3 J.G. Göhberg and M.G. Krejn, (1969). *Introduction to the theory of linear non selfadjoint operators*. English translation, Trans. Math. Monographs, **18**, Amer. Soc. Providence, 36-3157.

4 H. Hotteling, (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psycho.*, **24**.

5 M. Okamoto, (1969). *Optimality of principal components*, In Multivariate Analysis II, 673-685, Academic Press New-York.

6 A. Pousse et J.J. Téchené, (1997). Quelques propriétés extrémales des valeurs singulières d'un opérateur I, *Prob. and Math. Statist.*, **17**, 197-221.

7 A. Pousse et J.J. Téchené, (1997). Quelques propriétés extrémales des valeurs singulières d'un opérateur II, *Prob. and Math. Statist.*, **18**, 1-18.

8 C.R. Rao, (1964). The use and interpretation of PCA in applied research. *Sankhya A*, **26**, 329-358.

9 C.R. Rao, (1979). Separation theorems for singular values of matrices and their applications in multivariate analysis. *J. Multiv. Anal.*, 362-379.

10 H. Wielandt, (1955). An extremum property of sums of eigenvalues. *Proc. Amer. Math. Soc.*, **6**, 106-110.

.

Liste et adresse des intervenants et des participants

A. Baillo Univ. Carlos 3, Madrid, abaillo@est-econ.uc3m.es

T. Benchikh Univ. Sidi bel Abbes, tbenchikh@univ-sba.dz

J. Bigot IMAG, Grenoble, bigot@imag.fr

A. Boudou Univ. Toulouse 3, boudou@cict.fr

N. Bru Univ. Grenoble 2, noelle.bru@upmf-grenoble.fr

H. Cardot INRA Toulouse, cardot@toulouse.inra.fr

V. Couallier Univ. Bordeaux 2, couallier@sm.u-bordeaux2.fr

C. Crambes Univ. Toulouse 3, crambes.christophe@wanadoo.fr

M. Delecroix ENSAI, Rennes, delecroi@ensai.fr

S. Dossou-Gbete Univ. Pau, simplice.dossou-gbete@univ-pau.fr

R. Faivre INRA, Toulouse, faivre@toulouse.inra.fr

F. Ferraty Univ. Toulouse 2, ferraty@univ-tlse2.fr

J. Fine Univ. Toulouse 3, fine@cict.fr

A. Goïa Univ. Novara, Italie, goia@cict.fr

M. Goulard INRA, Toulouse, goulard@toulouse.inra.fr

J. Johannes Univ. Toulouse 1, jan.johannes@web.de

R. Lafosse Univ. Toulouse 3, lafosse@cict.fr

P. Loup IURC, Montpellier, p-loup@iurc.montp.inserm.fr

A. Mas Univ. Toulouse 3, mas@cict.fr

E. Mazza Univ. Toulouse 3, mazza@cict.fr

A. Nadja Univ. Sidi bel Abbes, n_azzedine@yahoo.fr

E. Molinari Univ. Montpellier 1, molinari@helios.ensam.inra.fr

L. Prchal Univ. Charles, Prague, lpsoft@centrum.cz

A. Rabhi Univ. Sidi bel Abbes, Rabhi-abbes@yahoo.fr

M. Rachdi Univ. Grenoble 2, Mustapha.rachdi@upmf-grenoble.fr

Y. Romain Univ. Toulouse 3, romain@cict.fr

P. Saint Pierre IURC, Montpellier, stpierre@iurc.montp.inserm.fr

P. Sarda Univ. Toulouse 2, sarda@cict.fr

J.J. Téchené Univ. Pau, jean-jacques.techene@univ-pau.fr

P. Vieu Univ. Toulouse 3, vieu@cict.fr

S. Viguier-Pla Univ. Perpignan, viguier@cict.fr

A.F. Yao Univ. Marseille, yao@com.univ-mrs.fr

E. Youndjé Univ. Rouen, Elie.youndje@univ-rouen.fr

A. Yousfate Univ. Sidi bel Abbes, yousfate_a@yahoo.com