INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet IS2

Inférence statistique pour l'industrie et la santé

Rhône-Alpes

THÈME 4A

Table des matières

1 Composition de l'équipe

Responsable scientifique

Gilles Celeux [DR Inria]

Personnel Inria

Florence Forbes [CR Inria]

Paulo Gonçalvès [CR Inria]

Anne Guérin-Dugué [maitre de conférence à l'INPG, détachée CR Inria, jusqu'au 31/08/01]

Stéphane Girard [maitre de conférences à l'université Montpellier II, détaché à l'Inria Rhône-Alpes depuis le 01/09/01]

Personnel des établissements partenaires

Christian Lavergne [professeur, université Paul Valéry, Montpellier]

Claudine Robert [professeur, université Joseph Fourier, Grenoble 1]

Chercheurs post-doctorants

Cyril Goutte [boursier Inria jusqu'au 30/09/01]

Gérard Boudjema [boursier Inria depuis le 01/10/01]

Yann Vernaz [ingénieur expert jusqu'au 30/09/01]

Chercheurs doctorants

Henri Bertholon [enseignant CNAM, à l'Inria jusqu'au 30/10/01]

Isabel Brito [enseignante détachée de l'université de Lisbonne]

Franck Corset [boursier Inria]

Cécile Delhumeau [CHU de Grenoble]

Jean-Baptiste Durand [boursier MESR]

Myriam Garrido [boursière Inria]

Olivier Martin [boursier MESR]

Nathalie Peyrard [boursière MESR jusqu'au 30/10/01]

Guillaume Bouchard [boursier Inria depuis le 01/10/01]

Julien Jacques [boursier Inria depuis le 01/11/01]

Stagiaire longue durée

Julien Gosme [université de technologie de Troyes]

Collaborateurs extérieurs

Christine Cans [médecin, association Rheops]

Jean Diebolt [DR CNRS université de Marne-la-Vallée]

Anatoli Iouditski [professeur, université Joseph Fourier, Grenoble 1]

Ollivier Taramasco [professeur, INPG]

Assistante de projet

Françoise de Coninck

2 Présentation et objectifs généraux

Le projet IS2 effectue des recherches en modélisation statistique. Plus spécifiquement, nous nous intéressons à la modélisation, à l'identification des modèles obtenus et à leur validation pour des systèmes ou des situations complexes pouvant intervenir dans le domaine industriel

ou biomédical.

IS2 s'intéresse essentiellement aux modèles, dits à structure de données incomplètes, où intrinsèquement une partie de l'information nécessaire à l'identification du phénomène étudié est manquante. Ces modèles sont courants (durées de vie censurées, modèles hétéroscédastiques[1], images dégradées, ...) et puissants (modèles à structure cachée, ...). Ils apparaissent dans de nombreux problèmes statistiques qui se posent en milieu biomédical et en milieu industriel. Ces modèles à observation partielle sont difficiles à estimer, de par leur nature intrinsèque et aussi parce qu'ils concernent eux-mêmes des systèmes complexes (montages industriels compliqués, existence d'une structure de dépendance temporelle ou spatiale, nombreuses variables en jeu,...). De ce fait, ces modèles sont en général faiblement identifiables en ce sens que, au vu des observations effectivement recueillies, plusieurs jeux différents de paramètres peuvent apparaître également bons. Cela se traduit par une multiplicité des *extrema* locaux des fonctions de contraste utilisées pour procéder à l'identification (vraisemblance, probabilité a posteriori,...). Ainsi, ces modèles requièrent une grande rigueur conceptuelle et méthodologique, le recours raisonné à un principe de parcimonie (retenir le modèle le moins complexe pour une qualité d'ajustement acceptable), et l'utilisation d'outils algorithmiques sophistiqués.

L'un des objectifs du projet IS2 est de proposer des méthodes efficaces d'estimation et d'évaluation de ces modèles. Pour l'estimation, nous privilégions les algorithmes dans lesquels les données manquantes sont restaurées par simulation ainsi que des algorithmes d'approximation stochastique pour l'estimation adaptative dans un cadre non paramétrique. La validation des modèles construits et identifiés est un élément important de notre recherche. Nous l'abordons par des tests statistiques ou, dans une perspective bayésienne, par le calcul de critères de parcimonie.

Les modèles considérés par IS2 sont souvent dictés par les problèmes qui nous sont soumis. Ainsi le choix de modèles bayésiens pour des problèmes d'analyse de défaillance s'explique-t-il par l'existence effective d'informations *a priori* et par la rareté des données de retour d'expérience. Dans le même ordre d'idée, notre intérêt pour la modélisation des événements rares et pour la prise en compte et la quantification d'opinions de plusieurs experts vient de problèmes qui nous ont été soumis par EDF. Les modèles hétéroscédastiques sont eux issus de problèmes concrets dans les domaines de la sélection en génétique, le contrôle de production ou l'analyse de séries financières.

L'inverse est vrai également. C'est donc notre culture sur les modèles à structure cachée qui nous a conduits à nous intéresser au modèle de champ de Markov caché pour l'analyse statistique d'image.

3 Fondements scientifiques

3.1 Modéles à structure cachée

Participants : Isabel Brito, Gilles Celeux, Jean-Baptiste Durand, Florence Forbes,

[1] On appelle modèle hétéroscédastique un modèle qui introduit une modélisation spécifique de la variance à l'aide de variables explicatives.

Nathalie Peyrard, Paulo Gonçalvès.

Mots clés : données manquantes, mélange de lois, algorithme EM, algorithme stochastique, combinaison et choix de modèles, analyse discriminante, analyse d'image, champ de Markov caché, analyse bayésienne.

Résumé : *Les modèles à structure cachée constituent un domaine important de la statistique aussi bien par leurs applications (classification, analyse du signal ou de l'image) que par les problèmes algorithmiques et théoriques (choix de modèles notamment) qu'ils soulèvent. L'analyse statistique d'image est un domaine relevant de ce type de modèles. Nous détaillons plus particulièrement le modèle de champ de Markov caché utilisé en analyse d'image.*

3.1.1 Généralités

Le projet IS2 s'intéresse à des modèles statistiques paramétriques, θ étant le paramètre à estimer, où les données complètes $x = x_1, \ldots, x_n$ se décomposent de manière naturelle en données observées $y = y_1, \ldots, y_n$ et en données manquantes $z = z_1, \ldots, z_n$. Les données manquantes z_i représentent l'appartenance à une catégorie d'objets parmi K. La densité des données complètes $f(x \mid \theta)$ et celle des données observées $f(y \mid \theta)$ sont liées par la relation $f(y \mid \theta) = \int f(x \mid \theta) dz = \int f(y, z \mid \theta) dz$. La loi marginale d'une donnée observée s'écrit comme un mélange fini de lois,

$$f(y_i \mid \theta) = \sum_{k=1}^{K} P(z_i = k \mid \theta) f(y_i \mid z_i, \theta) \, .$$

Un tel modèle peut par exemple être utilisé pour rendre compte des variations de la taille des adultes. Une variable cachée (le sexe) explique entièrement les variations entre les tailles, les variations de taille pour les personnes de même sexe étant considérées comme la réalisation d'un bruit gaussien. Ce type de modèle à données incomplètes est intéressant car il est susceptible de mettre en évidence une variable discrète cachée qui explique l'essentiel des variations et par rapport à laquelle les données observées sont *conditionnellement* indépendantes. Les modèles de mélange de lois lorsque les z_i sont indépendants constituent une approche de plus en plus répandue en classification. Les modèles de chaîne de Markov cachée (resp. champ de Markov caché) correspondent au cas où les z_i sont les réalisations d'une chaîne (resp. champ) de Markov. Ils sont très utilisés en traitement du signal (reconnaissance de la parole, analyse de séquences génomiques, etc.) et de l'image (voir section 3.1.2).

Les algorithmes Du point de vue mathématique, ces modèles sont souvent difficiles à estimer du fait même de l'existence de données manquantes. Ils ont donné naissance à de nombreux algorithmes, dont le dénominateur commun est la restauration des données manquantes, mais qui diffèrent par leur stratégie de restauration. L'algorithme le plus utilisé est l'algorithme EM[MK97].

[MK97] G. McLachlan, T. Krishnam, *The EM algorithm and extensions*, John Wiley, New York, 1997.

Glossaire :

Algorithme EM C'est un algorithme très populaire pour l'estimation du maximum de vraisemblance de modèles à structure de données incomplètes. Chaque itération comporte deux étapes. L'étape E (*expectation*) qui consiste à calculer l'espérance conditionnelle de la vraisemblance des données complètes sachant les observations et l'étape M (*maximisation*) qui consiste à maximiser cette espérance conditionnelle.

Les versions stochastiques de l'algorithme EM, dont Gilles Celeux et Jean Diebolt comptent parmi les pionniers, incorporent une étape de simulation des données manquantes pour pouvoir travailler sur des données complétées.

Les algorithmes MCMC (*Markov Chain Monte Carlo*) sont définis dans un cadre bayésien. Partant d'une loi a priori pour les paramètres, ils simulent une chaîne de Markov, définie sur les valeurs possibles des paramètres, et qui a pour loi stationnaire la loi recherchée, à savoir la loi a posteriori des paramètres. À chaque étape, z est simulé selon sa loi conditionnelle courante sachant les observations.

L'étude du comportement pratique et des propriétés de ces algorithmes stochastiques constitue un thème de recherche traditionnel du projet.

Choix de modèles Un point important pour les modèles à structure cachée est le choix de la complexité du modèle et en particulier le choix du nombre K de catégories de la variable cachée. Dans ce domaine, très ouvert, de nombreuses approches sont en compétition et la stratégie adoptée dépend beaucoup du but poursuivi. Par exemple, dans un contexte de classification, l'objectif est surtout de restaurer les catégories manquantes z_i, alors que dans un contexte d'estimation de densités, il est plutôt d'estimer le paramètre θ. Cela étant, une approche répandue consiste à se placer dans un cadre bayésien non informatif et à chercher le modèle m qui maximise la vraisemblance intégrée[RW97]

$$f(y \mid m) = \int f(y \mid m, \theta)\pi(\theta \mid m)d\theta,$$

$\pi(\theta \mid m)$ étant une distribution de probabilité a priori non informative (c'est-à-dire ne favorisant pas de valeur particulière) du paramètre θ.

Analyse discriminante Dans un cadre décisionnel, on dispose d'un échantillon d'apprentissage étiqueté, c'est-à-dire d'un échantillon complet $x = (y, z)$. Le problème est alors de construire une règle de décision pour classer de futures unités pour lesquelles seules les valeurs y_i seront observées. Il s'agit alors d'un problème d'analyse discriminante, courant en diagnostic médical, ou en reconnaissance statistique des formes. Dans ce domaine, bien établi[McL92], de nombreuses méthodes existent. La recherche consiste surtout, à l'heure actuelle, à proposer des techniques répondant à des contextes particuliers et à proposer des méthodes fiables lorsque les échantillons d'apprentissage sont de faible taille. C'est ce dernier point que nous privilégions dans notre recherche.

[RW97] K. Roeder, L. Wasserman, « Practical Bayesian density estimation using mixtures of normals », *Journal of the American Statistical Association 92*, 1997, p. 894–902.

[McL92] G. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 1992.

3.1.2 La modélisation statistique en analyse d'image

Les modèles à structure cachée apparaissent naturellement en analyse d'image où les phénomènes aléatoires ont un rôle important. Les données mises en jeu sont spatialement localisées et induisent l'utilisation de modèles probabilistes spatiaux. Ceux-ci soulèvent de nombreuses questions de modélisation et d'inférence statistique et n'ont cessé de gagner de l'intérêt. En particulier, le choix de modèles appropriés et l'estimation des paramètres associés aux modèles utilisés sont des questions essentielles pour aller vers une automatisation des algorithmes et tirer tout le profit de la richesse des modèles stochastiques. Ces problèmes, abondamment traités, restent cependant ouverts. En effet, un effort d'ordre méthodologique (recherche d'estimateurs précis et robustes) et d'ordre algorithmique (réduction des temps de calcul) reste à faire.

Segmentation et restauration d'image Des mécanismes de dégradation des observations sont souvent inhérents aux problèmes d'images. Dans les problèmes de segmentation, de classification ou de restauration d'image, il s'agit de construire ou de retrouver une image inconnue z lorsque seule une version dégradée y est observée. Cela relève naturellement des modèles à structure cachée. Les images sont constituées d'un ensemble S de pixels qui peuvent prendre une valeur parmi un petit nombre K de couleurs non ordonnées (les classes). Dans la suite nous noterons z_i (resp. y_i) la valeur de l'image z (resp. y) au pixel i et plus généralement z_A (resp. y_A) la restriction de z (resp. y) à un sous-ensemble A de pixels.

Une approche possible, bien fondée statistiquement, est l'analyse d'image dite bayésienne. Elle fournit des solutions élégantes et a connu des développements considérables depuis des premiers travaux tels que ceux de D. et S. Geman[GG84] ou Besag[Bes86]. L'intérêt de cette approche est la possibilité d'introduire explicitement des connaissances a priori, notamment sur la structure spatiale des images analysées, dans la modélisation des mécanismes de dégradation des données. Elle a aussi l'avantage de fournir un cadre général dans lequel une grande variété d'applications peuvent être envisagées, par exemple en imagerie médicale et satellitaire, sismologie, astronomie, etc.

Dans cette approche, le processus physique d'acquisition des données est pris en compte à travers une vraisemblance $f(y \mid z, \theta)$ qui précise la probabilité d'observer des données y lorsque l'image non dégradée est z. Le paramètre θ est ici souvent interprété comme un paramètre de bruit. L'information sur la « vraie » image z est prise en compte à travers une loi de probabilité, $f(z \mid \beta)$, fixée en fonction du problème traité et qui peut dépendre d'un paramètre β, réglant, par exemple, le niveau des dépendances spatiales. Dans ce modèle, une source d'information importante est la loi conditionnelle de z sachant les observations y, donnée par la formule de Bayes suivante

$$f(z \mid y, \theta, \beta) \quad \propto \quad f(y \mid z, \theta) f(z \mid \beta) \ . \tag{1}$$

Elle gère la probabilité que la vraie image soit z sachant que l'image dégradée observée est y. Un candidat naturel pour z est la valeur qui maximise $f(z \mid y, \theta, \beta)$, encore appelée MAP pour

[GG84] S. Geman, D. Geman, « Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images », *I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence 6*, 1984, p. 721–741.

[Bes86] J. Besag, « On the statistical analysis of dirty pictures », *Journal of the Royal Statistical Society, series B 48*, 1986, p. 259–302.

maximum a posteriori. Une autre possibilité est l'estimateur MPM (*marginal posterior mode*) obtenu en maximisant individuellement les probabilités marginales a posteriori, $f(z_i \mid y, \theta, \beta)$. Cela revient à maximiser le nombre moyen de pixels bien classés. D'autres possibilités existent, que nous ne mentionnons pas ici.

Lorsque les paramètres θ et β sont connus, la loi conditionnelle (1) peut être simulée à l'aide d'un échantillonneur de Gibbs[GG84] en considérant chaque pixel successivement. Lorsque l'on se trouve au pixel i, la valeur en ce site est remplacée par une valeur tirée au hasard suivant la loi conditionnelle $f(z_i \mid z_{S\setminus\{i\}}, y, \theta, \beta)$. En couplant cette technique avec un principe de recuit simulé, D. et S. Geman[GG84] ont proposé une méthode pour rechercher le MAP dans les cas où une énumération directe est impossible. L'échantillonneur de Gibbs peut également être utilisé pour appliquer la règle du MPM en calculant des probabilités empiriques d'appartenance de chaque pixel à une classe. De telles approches rencontrent les problèmes usuels de convergence des algorithmes de type MCMC et sont généralement lentes. Les solutions fournies peuvent être sensibles aux propriétés globales non réalistes des modèles adoptés. Une alternative plus rapide, et qui repose sur des propriétés locales des modèles sous-jacents, est l'algorithme déterministe ICM[Bes86]. La convergence n'est toutefois garantie que vers un maximum local de (1) et l'algorithme peut être très sensible aux conditions initiales. À partir d'une image initiale $z^{(0)}$, à l'itération $t+1$, un pixel i est choisi et sa valeur est mise à jour en lui donnant la valeur qui maximise $f(z_i \mid z_{S\setminus\{i\}}, y, \theta, \beta)$.

Modélisation markovienne L'approche bayésienne nécessite la spécification de la distribution $f(z \mid \beta)$. Il s'agit essentiellement de modéliser des phénomènes ou des contraintes physiques sous-jacentes. En particulier, il est raisonnable de supposer que des pixels voisins ont plus de similarités que des pixels éloignés. De telles caractéristiques locales peuvent être prises en compte à travers les probabilités conditionnelles qu'un pixel i prenne la valeur z_i connaissant la valeur de tous les autres pixels $z_{S\setminus\{i\}}$. Les champs de Markov sont des modèles dans lesquels la dépendance est réduite aux pixels dans un proche voisinage de i. Ils permettent donc de prendre en compte les dépendances spatiales entre les pixels d'une image mais ceci au prix de calculs importants. En particulier, lorsque le paramètre β du modèle est inconnu, son estimation est un problème ouvert.

Algorithmes non supervisés Les méthodes indiquées ci-dessus supposent les paramètres θ et β connus. En pratique, ces paramètres doivent être estimés à partir des informations disponibles, ce qui peut présenter certaines difficultés dans le cas des modèles markoviens. Lorsque l'on dispose de données pour lesquelles on connaît à la fois les observations y et la vraie image z, on peut envisager d'estimer les paramètres β et θ lors d'une phase d'apprentissage. Très souvent, de telles données ne sont pas disponibles. Il arrive également que la phase d'apprentissage demande l'intervention d'un opérateur humain dans des situations où une automatisation du système est souhaitée. Ainsi, la recherche d'algorithmes non supervisés est-elle d'un grand in-

[GG84] S. Geman, D. Geman, « Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images », *I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence 6*, 1984, p. 721–741.

[Bes86] J. Besag, « On the statistical analysis of dirty pictures », *Journal of the Royal Statistical Society, series B 48*, 1986, p. 259–302.

térêt pratique. Dans le cas le plus général, seules les données y sont observées et z, θ, β sont inconnus. Pour appliquer les méthodes précédentes, les paramètres doivent donc être estimés en même temps que l'image z.

Notons que plusieurs problèmes peuvent être envisagés. Il peut s'agir d'estimer seulement θ et β. C'est le cas lorsque l'on souhaite faire de la sélection de modèles sur des observations bruitées, ou plus généralement estimer des paramètres dans des problèmes à données manquantes. Il peut également s'agir d'estimer seulement z, par exemple dans des situations de classification ou segmentation d'image. Beaucoup des algorithmes fournissent à la fois des estimations de z et des paramètres θ et β de sorte que la distinction précédente peut sembler inutile. Nous décrivons toutefois dans [6] un algorithme fournissant une segmentation z sans donner une estimation précise de β, ce qui permet d'éviter des calculs coûteux.

3.1.3 Dépendance markovienne multi-échelle sur les coefficients d'ondelette

Les décompositions en ondelettes (orthogonales) fournissent pour une large classe de signaux une représentation *parcimonieuse*, dans laquelle peu de coefficients ont une amplitude significativement non nulle. Bien que ces décompositions ne génèrent pas *stricto sensu* une base de Kharunen-Loeve pour les processus étudiés, il est raisonnable dans une majorité de cas de négliger les corrélations résiduelles entre coefficients. Ici, nous nous intéressons à des situations où, précisément, il est important de ne pas sous-estimer ces corrélations. C'est le cas notamment des processus structurés en échelle, terminologie intentionnellement vague pouvant désigner les processus à mémoire longue, aussi bien que des signaux présentant des couplages entre plusieurs modes spectraux (par exemple des modes harmoniques). Nous proposons alors de modéliser ces interactions par des dépendances markoviennes sur des états cachés des coefficients d'ondelette structurés selon un arbre diadique multirésolution.

Le modèle statistique ainsi défini sur les coefficients d'ondelette est un modèle à structure cachée pour lequel existent des algorithmes de calcul et de maximisation de la vraisemblance comparables à l'algorithme avant-arrière pour les chaînes de Markov cachées.

Ainsi, si l'on privilégie l'axe temporel, on s'attache à modéliser la dépendance statistique de l'état d'un système conditionnellement à son passé relatif à une échelle de temps (caractéristique) donnée. Si, en revanche, on privilégie l'axe des échelles, on vise à caractériser les interactions entre les différents modes spectraux (ou échelles de temps). On peut ainsi envisager de repérer grâce à ces modèles, des comportements en loi d'échelle (auto-similarité globale ou locale, longue dépendance), ou, ce qui nous intéresse davantage, des transitions dans cette dynamique d'échelle (processus multi-échelle, scalings non stationnaires...).

3.2 Modèles linéaires généralisés et hétéroscédasticité

Participant : Christian Lavergne.

Mots clés : modèle linéaire généralisé, hétéroscédasticité, structure exponentielle, modèle à effets aléatoires, modèle ARCH.

Résumé : *La régression a pour objet la modélisation et l'étude de la relation entre une variable dite réponse et une ou plusieurs autres variables dites explicatives ou*

*régresseurs. Dans ce cadre, choisir un estimateur revient à minimiser une distance entre un modèle et des observations. À la base, il y a la régression linéaire et la méthode des moindres carrés. Cette notion, connue de tout statisticien, s'appuie sur trois hypothèses fondamentales. La première est le lien linéaire qui existe entre la variable réponse et les variables explicatives. La deuxième réside dans la loi de probabilité des erreurs supposée gaussienne. La troisième est l'*homoscédasticité *du modèle : la variance des observations est indépendante des variables explicatives. Afin de relâcher deux des hypothèses fortes de la régression linéaire, la loi des erreurs et l'homoscédasticité, diverses théories se sont développées en parallèle.*

Nous donnons ici la définition de plusieurs types de modèles généralisant le modèle linéaire et qui font l'objet de recherches dans le projet IS2.

Les modèles linéaires mixtes Un modèle linéaire mixte (L2M) est défini par la donnée d'un vecteur aléatoire Y de dimension n :

$$Y = X\beta + U\xi + \epsilon,$$

U étant une matrice connue de dimension $n \times q$ fixée et ξ un vecteur aléatoire de $\mathbf{R}^q$ non observé. Les distributions des variables aléatoires ξ et ϵ sont supposées gaussiennes. La matrice X $n \times p$ de rang p est connue, et le vecteur p-dimensionnel β ainsi que les variances de ξ et ϵ sont les paramètres inconnus du modèle.

Les modèles linéaires généralisés Un modèle linéaire généralisé (GLM) est défini par la donnée :

i) d'un vecteur aléatoire Y de dimension n ayant des composantes indépendantes et dont la fonction de vraisemblance pour une réalisation $y = (y_1, \ldots, y_n)$ s'écrit :

$$L_y(\theta, \phi) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}, \tag{2}$$

où a, b et c sont des fonctions réelles données et θ le paramètre d'intérêt.

ii) d'un prédicteur linéaire η $(\eta_i)_{i=1,\ldots,n} relil'esprancemathmatique \mathrm{E(Y)} = \mu$ par une fonction g : $\eta = g(\mu)$, la fonction g étant la *fonction de lien* du modèle.

Le prédicteur linéaire η est défini dans le cas d'un GLM par la donnée d'une matrice X de dimension $n \times p$, de rang p, appelée matrice du plan d'expérience, et d'un vecteur p-dimensionnel β, paramètre inconnu du modèle, tel que $\eta = X\beta$.

Les modèles ARCH (auto-régressifs conditionnellement hétéroscédastiques) Un processus stochastique réel $\varepsilon_t, t \in Z$ est dit ARCH(p) s'il est défini par une équation du type :

$$\varepsilon_t = u_t h_t \text{ avec } h_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2$$

où α_i est un paramètre inconnu positif pour $i = 0, \ldots, p$ et $(u_t)_{t \in Z}$ est une suite de variables aléatoires à valeurs réelles, indépendantes, équidistribuées, de moyenne nulle et de variance un.

On appelle modèle à erreur ARCH un modèle de la forme :

$$y_t = \mu_t(\theta) + \varepsilon_t \text{ où } \varepsilon_t \text{ est un processus ARCH,}$$

et $\theta \in \mathbf{R}^k$ est un paramètre inconnu.

Les modèles linéaires généralisés mixtes Un mixte GL2M est défini par la donnée d'un vecteur de réponse y et d'une composante aléatoire ξ de $\mathbf{R}^q$ non observée, telle que la vraisemblance conditionnelle de y sachant ξ soit celle d'un GLM avec comme prédicteur linéaire :

$$\eta_\xi = X\beta + U\xi,$$

U étant une matrice de dimension $n \times q$ fixée. La distribution de la variable ξ est supposée gaussienne.

Les modèles GLM-ARCH Un modèle GLM-ARCH d'ordre q est défini par la donnée d'un vecteur de réponse $y = (y_1, \ldots, y_t, \ldots, y_T)$ et d'une suite de prédicteurs aléatoires :

$$\eta_t = (X\beta_0)_t + \beta_1 g(Y_{t-1}) + \beta_2 g(Y_{t-2}) + \ldots + \beta_q g(Y_{t-q}) \text{ pour } t > q,$$

les valeurs initiales $\eta_1, \ldots, \eta_q$ étant fixées, de sorte que la vraisemblance conditionnelle de y sachant le passé soit celle d'un GLM avec comme prédicteur linéaire η_t.

3.3 Estimation de lois d'échelle par ondelettes

Participant : Paulo Gonçalvès.

Mots clés : estimation, lois d'échelle, ondelettes, spectres de singularités.

Résumé : *L'efficacité des décompositions en ondelettes pour caractériser les comportements en loi d'échelle des signaux ou des processus est maintenant largement établie. Dans le cas de processus aléatoires, nous nous intéressons aux performances statistiques des estimateurs empiriques des exposants d'échelle (ou de singularité) construits à partir des coefficients d'ondelette.*

Ce travail a été effectué en collaboration avec Rudolf Riedi (Rice University, Houston (TX), USA). Soit $x(t)$ la trajectoire d'un processus aléatoire. La régularité hölderienne locale de $x(t)$ est définie par

$$\alpha(t) := \limsup_{\varepsilon \to 0} \frac{1}{\log_2(2\varepsilon)} \log_2 \sup_{|s-t|<\varepsilon} |x(s) - x(t)|.$$

Le spectre de singularités de Hausdorff permet de mesurer géométriquement la distribution des régularités $\alpha(t)$, selon

$$d(\alpha) = \dim_{\mathcal{H}}\{t : \alpha(t) = \alpha)\},$$

où $\dim_{\mathcal{H}}\{E\}$ désigne la dimension de Hausdorff de l'ensemble E. En pratique cette définition se heurte à plusieurs obstructions. D'une part il n'est pas réaliste d'espérer accéder en chaque

point t de la trajectoire de x à la régularité hölderienne $\alpha(t)$. D'autre part, on ne sait pas calculer l'infinité de dimensions de Hausdorff correspondant à chacune des valeurs de α.

Le *formalisme multifractal* permet alors dans certains cas de substituer au spectre de Hausdorff un spectre qui lui est égal, mais qui est plus simple à estimer. Ce spectre, dit de Legendre, initialement défini sur les moments d'ordres supérieurs des accroissements du processus $\delta^{-1}|x(t+\delta) - x(t)|$, admet une formulation équivalente sur les coefficients en ondelette $\{C_{n,k}\}_{(n,k)\in\mathbb{Z}\times\mathbb{Z}}$ issus de la décomposition de x

$$C_{n,k} := \int x(t)\, 2^{n/2} \psi^*(2^n t - k)\, \mathrm{d}t.$$

Comme pour l'analyse classique construite sur les accroissements du processus, l'estimateur empirique des moments d'ordre q des coefficients d'ondelette permet d'estimer *la fonction de structure* de x

$$S^n(q) := 2^{-n} \sum_{k=0}^{2^n-1} |C_{n,k}|^q.$$

Les différentes lois d'échelle qui composent le processus x se traduisent alors par une évolution linéaire de la fonction de structure selon l'échelle n dans un schéma bi-logarithmique. La pente de ces évolutions est donnée par *la fonction de partition* :

$$\tau(q) := \liminf_{n\to\infty} -\frac{1}{n} \log_2 S^n(q),$$

et *le spectre de Legendre* correspond simplement à la transformée de Legendre de celle-ci :

$$f(\alpha) \;:=\; \tau^*(\alpha) \;=\; \inf_{q\in\Re}(q\alpha - \tau(q)).$$

En toute généralité on a la relation $d(\alpha) \leq f(\alpha)$, entre spectre de Hausdorff et spectre de Legendre. Néanmoins, pour certains processus il y a égalité, et on dit alors qu'ils vérifient le formalisme multifractal.

Pour différentes classes de processus (mono- ou multi-échelles), nous nous intéressons à la caractérisation des performances statistiques de cet estimateur, et proposons des améliorations méthodologiques pour rendre l'estimation de $f(\alpha)$ fiable et robuste pour une classe de processus la plus large possible.

4 Domaines d'applications

4.1 Fiabilité industrielle

Participants : Henri Bertholon, Gilles Celeux, Franck Corset, Jean Diebolt, Cyril Goutte, Christian Lavergne, Myriam Garrido.

Un domaine d'applications important d'IS2 a trait à la sûreté de fonctionnement et à l'analyse de fiabilité de systèmes mécaniques. Il se concrétise dans le cadre de conventions d'étude et recherche (CERD) avec le groupe « retour d'expérience » et le département « Surveillance, Diagnostic, Maintenance » de l'EDF-DER. Les problèmes auxquels nous sommes confrontés relèvent

de l'analyse de durées de vie de systèmes non réparables pouvant être sujets à vieillissement, l'étude de la cinétique de dégradation de systèmes passifs (tuyaux par exemple) et la modélisation statististique de modes de défaillance prenant en compte l'avis d'experts. Les données dont nous disposons pour ces études viennent du retour d'expérience associé aux opérations de maintenance préventive. Elles sont alors de nature quantitative. Sinon il s'agit d'avis d'experts le plus souvent qualitatifs.

Les modèles de durée de vie ou d'occurrence d'incidents que nous proposons doivent prendre en compte la rareté des défaillances observées entrainant la présence largement majoritaire de données censurées.

Glossaire :
Durée de vie censurée Une durée de vie est censurée à droite si, sa valeur exacte étant inconnue, on sait seulement qu'elle est plus grande qu'une valeur appelée censure.

Dans bien des cas le nombre total de données est faible. Par ailleurs les systèmes mécaniques sont souvent sujets à vieillissement. Cela nous conduit à nous intéresser à des modèles paramétriques gouvernés par des lois de Weibull.

Glossaire :
Loi de Weibull Une durée de vie suit une loi de Weibull si sa densité s'écrit, pour $x > 0$,

$$f(x) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1} \exp\left(-\frac{x}{\eta}\right)^{\beta},$$

η est un paramètre d'échelle et β un paramètre de forme qui traduit le vieillissement ($\beta < 1$ défaut de jeunesse, $\beta = 1$ pas de vieillissement et $\beta > 1$ vieillissement).

Plus généralement, on est amené à modéliser des événements rares (fissures exceptionnelles, sollicitations extrêmes, ...). Ainsi, l'estimation de *quantiles extrêmes* est-il un sujet de recherche important de notre équipe. De plus, cela nous a incité à considérer la modélisation bayésienne, prenant en compte des informations a priori ne relevant pas du retour d'expérience, comme alternative à l'estimation par maximum de vraisemblance.

4.2 Statistique biomédicale

Participants : Christine Cans, Gilles Celeux, Cécile Delhumeau, Olivier Martin, Christian Lavergne, Claudine Robert.

Travail en collaboration avec Jérome Fauconnier du CHU de Grenoble.

Notre deuxième domaine d'intervention, moins développé, concerne les applications biomédicales. Les problèmes que nous considérons concernent surtout l'analyse de données hospitalières ou la détermination de facteurs de risque de maladies. Ils se concrétisent dans le cadre d'actions avec les collaborateurs extérieurs du projet, médecins au CHU de Grenoble, et membres du laboratoire TIMC de l'Imag. Nous sommes amenés à mettre en œuvre des modèles assez variés de type modèle linéaire et des techniques d'analyse multidimensionnelle des données (arbres d'induction, analyses factorielles).

Cette année, nous avons ouvert un nouveau sujet de recherche, l'analyse de données issues de puces à ADN (ou biopuces). Ce domaine connait un développement important en raison

des problèmes statistiques qui y sont liés et des résultats attendus en génomique fonctionnelle. À partir de mesures donnant le niveau d'expression de plusieurs milliers de gènes, il s'agit de déterminer leurs implications dans des processus biologiques. Notre travail actuel s'est axé sur les problèmes de normalisation et de recherche de gènes différentiellement exprimés. Dans la suite, nous souhaitons en tirer des éléments pour aborder la classification des profils d'expression.

5 Logiciels

5.1 Boîte à outils MATLAB de modélisation non linéaire

Participant : Anatoli Iouditski.

Mots clés : identification, modélisation « boîte-noire », Matlab toolbox.

En coopération avec Lennart Ljung et Peter Lidskog de l'université de Linköping, Qinghua Zhang et Bernard Delyon de l'Irisa, Rennes, nous préparons, depuis l'automne 1996, une boîte à outils Matlab. Cette boîte à outils est conçue comme une extension de la boîte à outils System Identification (SI-Toolbox) de Lennart Ljung, qui servira à la modélisation de systèmes dynamiques non linéaires. Les techniques utilisées sont les algorithmes adaptatifs d'estimation non paramétrique, les réseaux de neurones et les réseaux d'ondelettes. Les modèles proposés sont pour l'essentiel de type auto-régressif non linéaire avec quelques extensions spécifiques pour lesquelles on dispose de bons algorithmes. La boîte à outils sera distribuée par Mathworks.

En ce qui concerne les services offerts, ce sont des outils d'identification par des modèles de type régression/auto-régression non linéaires, des modèles de type Wiener et Hammerstein. L'originalité consiste en l'utilisation intensive d'algorithmes non itératifs d'estimation non paramétrique basés sur le triage adaptatif des estimées, *algorithmes d'arbre* ; ces algorithmes sont développés depuis quelques années dans le projet SIGMA2, et utilisent des polynômes locaux pour identifier des systèmes dont l'entrée est de dimension élevée. Ces méthodes ne font pas appel à la rétropropagation ni à des méthodes de gradient.

Étant complètement adaptatifs, ces algorithmes permettent de s'affranchir des réglages difficiles d'algorithmes. On gagne ainsi en qualité d'estimation de manière spectaculaire, et l'on évite les écueils liés à l'accrochage d'une méthode d'optimisation récursive sur un minimum local[J+95]. Outre les services d'identification proprement dite, on offre des moyens d'estimation/validation d'une modélisation restreinte :

- tests de linéairité de modèles ;
- modélisation par des fonctions "ridge" ;
- réduction de dimension de modèle et des tests correspondants.

[J+95] A. JUDITSKY *et al.*, « Nonlinear Black-Box Modelling in System Identification », *Automatica 31*, 12, 1995, p. 1725–1750.

5.2 Le logiciel MIXMOD

Participants : Gilles Celeux, Florent Langrognet[2], Yann Vernaz.

En collaboration avec Christophe Biernacki et Florent Langrognet de l'Université de Franche-Comté et Gérard Govaert de l'Université de Technologie de Compiègne, le projet IS2 a développé MIXMOD (Mixture Modelling), logiciel dédié à l'estimation de mélanges gaussiens. Les mélanges multivariés gaussiens constituent un modèle de référence en analyse discriminante et en classification, mais ausi en estimation semi-paramétrique de densités. MIXMOD propose un grand nombre de modèles autorisant des variations sur la forme, l'orientation, le volume et la taille des composants (ou classes) du mélange. L'estimation peut se faire par différents algorithmes (EM, EM stochastique et EM classification) qui peuvent être enchaînés pour de meilleures performances. Le choix des modèles peut se faire par différents critères (BIC, validation croisée, vraisemblance complétée intégrée, entropie) suivant l'objectif visé. Ce logiciel s'adresse aussi bien à un public expert qu'occasionnel par la possibilité de définir soi-même ses stratégies ou de s'en remettre à des choix par défaut.

MIXMOD a été écrit en C++ et des interface Scilab et Matlab ont été réalisés. Il est diffusé en *open source* et téléchargeable à l'adresse web suivante : `http://www.inrialpes.fr/is2/pub/software/MIXMOD`. Il sera prochainement diffusé sur le cdrom des logiciels libres distribués par l'Inria.

Deux évolutions essentielles sont prévues : l'extension à d'autres types de loi, comme les lois multinomiales intervenant dans le modèle des classes latentes pour la prise en compte de variables qualitatives, et l'extension vers des modèles de chaînes de Markov cachées. MIXMOD a été conçu pour être un système le plus ouvert possible avec la possibilité de définir de nouveaux modèles, de nouveaux critères et de nouveaux algorithmes comme également de nouveaux types de données.

5.3 Le projet SEL

Participant : Claudine Robert.

Travail en collaboration avec Marcos Perreau-Guimaraes et Bernard Ycart de l'équipe Prisme, université René Descartes.

Le portail web (`http://www.inrialpes.fr/is2/`) de statistique en ligne à l'usage des enseignants en mathématiques du secondaire a été diffusé à 30 000 exemplaires (tous les professeurs de mathématiques des lycées) et a déjà été demandé par divers pays. Un document d'accompagnement de 90 pages pour les nouveaux programmes des classes de première ainsi que diverses animations sur des parties des programmes ont été ajoutés.

Le portail SEL propose une initiation interactive à la statistique, articulée en trois couches.

- Une couche ARTICLES propose des textes, contenant des exemples d'utilisation de la statistique.
- La couche LEXIQUE contient un index des termes statistiques, référencés dans les articles et expliqués dans des pages séparées.

[2] université de Franche-Comté

- *Termes nodaux.* Ce sont des parties de termes simples ou développés plus précis. Par exemple « moyenne » renvoie à « moyenne empirique », « moyenne élaguée », « moyenne mobile ».
- *Termes simples.* Ils renvoient à une page contenant une brève définition, des liens vers les autres couches et un bouton cliquable « voir aussi » qui renvoie sur des termes proches.
- *Termes développés.* Ils renvoient à une page contenant le même type d'information que celle des termes simples, plus une applet illustrant le terme par une expérimentation interactive.

– La couche COURS est un cours de statistique au sens classique.

5.4 Le logiciel EXTREMES

Participants : Myriam Garrido, Jean Diebolt.

Dans le cadre de notre collaboration avec le groupe « Retour d'expérience » de EDF-DER, nous avons poursuivi la programmation d'un logiciel interactif en Matlab, interne à EDF, et intitulé EXTREMES. Nous avons modifié les programmes initiaux en fonction des remarques et demandes de EDF-DER. Ce logiciel permet maintenant de réaliser toute la procédure du test ET et l'alternative basée sur une approche bayésienne (cf. section 7.2). Les quatre procédures proposées sont :

- Un test d'adéquation classique,
- Un test d'adéquation de la loi exponentielle aux excès,
- Le test ET sous ses différentes versions (à n'appliquer que lorsque les excès suivent une loi exponentielle),
- Une procédure de régularisation bayésienne (à appliquer principalement lorsque les résultats du test classique et du test ET sont en contradiction).

D'autre part, nous sommes en train d'implémenter une prodédure d'estimation des paramètres d'une loi GPD par une méthode bayésienne. Les programmes correspondants vont être livrés à EDF-DER et seront inclus dans EXTREMES.

6 Résultats nouveaux

6.1 Modèles à structure cachée

6.1.1 Stratégies d'obtention du maximum de vraisemblance pour les mélanges

Participant : Gilles Celeux.

Travail en collaboration avec Christophe Biernacki de l'université de Franche-Comté et Gérard Govaert de l'université de technologie de Compiègne.

La fonction de vraisemblance pour un mélange multidimensionnel comporte de nombreux maxima locaux. L'obtention du maximum global est d'autant plus importante que ce maximum entre dans la composition de nombreux critères de choix de modèles ; mais c'est souvent un problème difficile. Forts des nombreux algorithmes présents dans MIXMOD (EM, EM stochas-

tique (SEM), EM classification (CEM)) et de la facilité de les combiner, nous avons exploré la capacité de stratégies simples pour accéder à cet optimum global [45]. Nous avons comparé des stratégies simples pour obtenir l'estimateur du maximum de vraisemblance d'un mélange par l'algorithme EM. Ces stratégies visent à bien initialiser l'algorithme EM. Elles sont fondées sur une initilisation au hasard, l'usage de versions classifiantes CEM ou stochastiques SEM de l'algorithme EM, ou encore sur l'utilisation préalable de courtes exécutions de l'algorithme EM lui-même. Nous les avons comparées dans le contexte des mélanges gaussiens multivariés par des expérimentations numériques sur des données simulées et réelles. Les principales conclusions sont les suivantes. L'initialisation au hasard qui est certainement la stratégie la plus répandue est souvent battue par les autres qui peuvent lui être préférées. De plus, il s'avère que la répétition d'exécutions des procédures est généralement bénéfique car une unique exécution peut souvent aboutir à une solution sous-optimale. Sinon, aucune des stratégies envisagées ici ne peut être considérée meilleure que les autres et il est difficile de cerner des situations où une stratégie particulière est censée se comporter mieux que les autres. Cependant, nous recommandons la stratégie qui consiste à initialiser l'algorithme EM par de courtes exécutions préalables de lui-même. Cette stratégie, qui est nouvelle, a plusieurs avantages. Elle est simple, marche souvent bien dans des situations très variées et semble peu sensible à des données bruitées.

6.1.2 Convergence de l'algorithme Monte-Carlo EM (MCEM)

Participant : Nathalie Peyrard.

L'algorithme MCEM est une version stochastique de l'algorithme EM dans laquelle la distribution conditionnelle des données cachées est approchée par une méthode de type Monte-Carlo. Dans le contexte de la segmentation markovienne d'image, le comportement de cet algorithme a été illustré de manière expérimentale, mais peu de résultats théoriques sont disponibles. Lorsque le paramètre spatial est connu, Comer et Delp ont établi des premiers résultats théoriques qui laissent espérer au mieux la convergence en probabilité de l'estimateur MCEM. Nous avons amélioré ces résultats en montrant que sous des conditions raisonnables de régularité, la suite des log-vraisemblances engendrée par l'algorithme augmente presque sûrement. Ceci permet de relier l'algorithme MCEM aux algorithmes EM généralisés et de définir des conditions suffisantes pour assurer que presque sûrement la suite des estimateurs MCEM converge vers le maximiseur de la vraisemblance du modèle [52].

6.1.3 Approximation du champ moyen et segmentation d'images

Participants : Gilles Celeux, Florence Forbes, Nathalie Peyrard.

Dans le cadre de la segmentation markovienne non supervisée, l'algorithme EM est une méthode classique pour l'estimation des paramètres puisqu'il s'agit d'un problème à données manquantes. Cependant, la complexité du modèle de champ de Markov rend impossible la mise en œuvre directe de cet algorithme. Une solution, assez coûteuse, consiste à utiliser des simulations de Monte-Carlo. Nous proposons des méthodes simples et moins lourdes, basées sur l'approximation en champ moyen d'un champ de Markov, issue de la mécanique statistique

[Cha87]. L'intérêt de cette méthode est de se ramener à un système de variables indépendantes, plus facile à manipuler.
Elle a déjà été utilisée par [Zha92] pour la mise en œuvre de l'algorithme EM et a montré des résutats prometteurs. Nous avons étudié, puis généralisé le principe du champ moyen dans ce contexte. Il en résulte une classe d'algorithmes, *les algorithmes* EM *de type champ moyen*, dont fait partie l'algorithme de Zhang, et qui présentent l'avantage de prendre en compte la structure markovienne tout en ayant une simplicité de mise en œuvre identique au cas indépendant. Parmi ces procédures, nous distinguons *l'algorithme en champ simulé*, proche de l'algorithme Monte-Carlo EM (MCEM), en raison de ses performances en termes d'estimation et de segmentation dans le cas d'images simulées et réelles. Nous mettons notamment en évidence les limites de l'utilisation de l'approximation en champ moyen pour la segmentation non supervisée, qui a tendance à fournir des segmentations trop lissées, phénomène que l'on évite avec l'algorithme en champ simulé. D'autre part cet algorithme hérite des bonnes caractéristiques de EM en ce sens qu'en travaillant avec des probabilités et non pas avec une restauration de type MAP il n'induit pas de biais sur l'estimation des paramètres.
Ces travaux ont été appliqués dans un premier temps pour un modèle de Potts simple. Dans le cadre du stage de maitrise de Franz Chouly (Ensimag), nous avons considéré un modèle plus complexe, le modèle de Potts avec champ externe. Les résultats obtenus confirment les conclusions précédentes.

6.1.4 Analyse statistique d'Images à Résonance Magnétique (IRM) pour la détection et l'identification de tumeurs

Participants : Florence Forbes, Nathalie Peyrard.

Travail en collaboration avec Chris Fraley et Adrian Raftery (Statistics Department, University of Washington, Seattle), David Goldhaber (Toshiba MRI inc., San Francisco) et Dianne Georgian-Smith (Harvard Medical School, Department of Radiology, Massachusetts General Hospital, Boston).

L'imagerie à résonance magnétique (IRM) apparaît comme un outil puissant d'aide au diagnostic des lésions du sein. C'est une technique sensible car les carcinomes réagissent fortement après injection d'un agent de contraste (Gadolinium) mais peu spécifique car des lésions bénignes peuvent également réagir fortement. Dans le but d'en améliorer la spécificité, nous proposons d'analyser les courbes d'évolution de l'absorption de l'agent de contraste. Une telle analyse nécessite la sélection préliminaire d'une région d'intérêt, souvent faite de manière subjective. Nous proposons ici une sélection automatique basée sur des techniques d'analyse de données multidimensionnelles et des méthodes statistiques de segmentation d'images. En plus du classique taux d'accroissement de l'intensité, nous utilisons d'autres paramètres pour grouper les régions du sein ayant des caractéristiques semblables puis nous sélectionnons la région qui correspond à un accroissement fort et rapide. Nous proposons ensuite des outils pour analyser les courbes dans la région sélectionnée en vue de déterminer si les lésions sont bénignes

[Cha87] D. Chandler, *Introduction to Modern Statistical Mechanics*, Oxford University Press, 1987.

[Zha92] J. Zhang, « The Mean Field Theory in EM Procedures for Markov Random Fields », *IEEE Transaction on signal processing 40*, 10, 1992, p. 2570–2583.

ou malignes. Nous appliquons cette procédure sur quelques cas dont le diagnostic est connu et obtenons des résultats prometteurs en accord avec la nature des lésions diagnostiquées [49].

Les techniques précédentes font surtout appel à des techniques de classification non supervisée mais nous avons aussi envisagé, dans le cadre du stage de DEA d'A. Regeasse [62], des techniques d'analyse discriminante pour l'analyse du signal dans les régions d'intérêt. Cette étude est restée à un stade préliminaire car nous ne disposions pas de données suffisantes pour réellement la valider. Précisons que de nouvelles données ont été acquises lors du séjour de F. Forbes à l'université de Washington à Seattle (cet été). Leur mise sous forme exploitable pour l'application de nos techniques est en cours. À cette occasion, nous avons également bénéficié de nouveaux contacts avec des experts, radiologues. Cette recherche fait l'objet d'un projet financé par le NIH (National Institut of Health), sous la direction d'A. Raftery.

6.1.5 Algorithmes d'inférence pour les arbres de Markov cachés

Participants : Jean-Baptiste Durand, Paulo Gonçalvès.

Travail en collaboration avec Yann Guédon (Cirad, Montpellier).

Dans le cadre de la modélisation des dépendances multi-échelle sur les coefficients d'ondelette par des arbres de Markov cachés (cf. paragraphe 3.1.3), de nouveaux algorithmes pour le calcul de la vraisemblance et sa maximisation ont été réalisés. Ils visent à surmonter les limitations numériques des algorithmes d'inférence classique *haut-bas* qui surviennent quand le nombre de variables aléatoires augmente. Ces algorithmes, décrits dans [48] sont basés sur l'emploi de probabilités conditionnelles au lieu de probabilités jointes.

D'autre part, le problème de la restauration des états cachés par la méthode du maximum a posteriori (MAP) pour ces modèles n'ayant pas de solution connue, nous avons développé l'algorithme du MAP pour le modèle des arbres de Markov cachés [48].

6.1.6 Modélisation de suites finies par des chaînes de Markov cachées

Participants : Gilles Celeux, Jean-Baptiste Durand.

Le modèle de chaînes de Markov cachées est fréquemment utilisé en reconnaissance statistique des formes, notamment en reconnaissance de parole ou de gestes. Comme pour les mélanges de loi, l'un des problèmes qui reste à résoudre concerne le choix du nombre d'états cachés. Nous avons entrepris de l'attaquer en utilisant une évaluation de la déviance du modèle par des techniques de validation croisée. Dans le principe, cela consiste à diviser plusieurs fois l'échantillon en deux parties de taille éventuellement inégale, puis à estimer les paramètres sur une partie et à calculer la vraisemblance du modèle sur l'autre. En répétant l'opération, on obtient ainsi une vraisemblance moyenne qui sert de critère de sélection. Du fait de la dépendance markovienne, le découpage en deux parties n'est pas ici une opération anodine. Dans le cas où les deux parties sont tirées au hasard, cela nous a amené à adapter l'algorithme de Baum-Welch de calcul de l'estimateur du maximum de vraisemblance dans une chaîne de Markov cachée à observations manquantes. Dans le cas où la chaîne est divisée suivant la parité des indices, nous avons montré que les processus obtenus sont encore des chaînes de Markov cachées, ce qui permet d'utiliser l'algorithme de Baum-Welch pour l'estimation des

paramètres [36]. Les expérimentations menées sont encourageantes et semblent indiquer une certaine supériorité de la procédure se fondant sur un partionnement alternatif et équilibré de l'échantillon. Nous explorons actuellement le positionnement de la validation croisée par rapport au critère BIC ainsi que son interprétation théorique.

6.1.7 Modèles de chaînes de Markov cachées pour le suivi de contours

Participant : Gilles Celeux.

Travail en collaboration avec Jorge Marques et Jacinto Nascimento (ISR-IST, Lisbonne).

Nous avons développé un modèle de chaîne de Markov cachée et un algorithme pour son estimation par le maximum de vraisemblance pour un problème de suivi de lèvres (*lip tracking*). L'intérêt et la difficulté viennent de ce que le signal observé est lui-même régi par un modèle autorégressif. Les paramètres de ce processus autorégressif varient en fonction des états cachés qui eux sont la réalisation d'une chaîne de Markov homogène. L'algorithme a été écrit et programmé. Des expérimentations sur des données réelles et simulées sont en cours. Cette recherche s'effectue dans le cadre de la collaboration INRIA/ICCTI (Portugal).

6.1.8 Étude d'événements en finance

Participants : Christian Lavergne, Ollivier Taramasco.

Nous avons mené une recherche portant sur les «études d'événements» en finance. À partir d'un échantillon de sociétés ayant toutes subi le même évènement (par exemple, elles ont toutes été la cible d'une O.P.A. hostile, elles ont toutes émis le même type d'obligation convertible pour financer un projet ...), et décrites par l'historique de ces rentabilités boursières, nous proposons une méthode qui permet d'une part de déterminer pendant combien de séances consécutives l'événement a un impact sur les cours (le début de cette fenêtre est inconnu et elle peut être de longueur nulle) et d'autre part d'indiquer comment cet événement a changé la nature des distributions des rentabilités. Bien entendu, cette recherche repose sur un grand nombre d'hypothèses concernant la nature de la distribution des rentabilités avant et après l'événement ainsi que sur la façon dont l'événement agit sur les cours. Le modèle proposé est un modèle à chaîne de Markov cachée ; la méthode d'estimation est basée sur l'algorithme EM. Le choix entre les différents modèles proposés est envisagé à l'aide du critère BIC.

Le modèle à structure cachée pour l'étude d'événements ayant donné des résultats prometteurs, nous nous proposons d'étendre cette formalisation pour définir de nouvelles stratégies d'investissement en Bourse, en considérant que certaines variations des cours sont des réactions à des événements cachés.

6.1.9 Modélisation statistique de la chrominance pour l'indexation d'images

Participant : Anne Guérin-Dugué.

Travail en collaboration avec Catherine Berrut (Université Joseph Fourier), Christophe Biernacki (Université de Franche Comté) et Jeanny Hérault (Université Joseph Fourier).

Dans les applications d'indexation d'images, domaine de recherche en plein développement, la modélisation statistique des distributions de caractéristiques prend une part importante. Le résultat de la modélisation est utilisé comme index pour rechercher les images dans les bases de données, en estimant des différences entre modèles comme mesure de disimilarité. Dans ce cadre, l'information de chrominance prend une part primordiale. Nous avons poursuivi le travail débuté l'année dernière, en collaboration avec Christophe Biernacki et Jeanny Hérault [41].
Cette année, l'effort s'est focalisé sur l'implantation logicielle d'une plate-forme de recherche d'images par le contenu en utilisant les résultats de modélisation de la chrominance (associée ou non à la forme spatiale des régions) par mélange de fonctions gaussiennes. La plate-forme a été développée en JAVA-C++ , elle utilise la bibliothèque MIXMOD.
Les deux points actuellement à l'étude sont (i) la prise en compte de l'information spatiale avec une précision variable dans la distribution chromatique, et (ii) la mise en place d'un modèle statistique de reformulation de la requête d'image.

6.2 Choix de modèles en discrimination et classification automatique

6.2.1 Sélection de modèle pour les champs de Markov caché

Participants : Gilles Celeux, Florence Forbes, Nathalie Peyrard.

Travail en collaboration avec Adrian Raftery (Statistics Department, University of Washington, Seattle).

Dans de nombreuses situations en analyse d'images le nombre de classes codant la segmentation est inconnu. D'autres caractéristiques d'un modèle markovien doivent parfois aussi être choisies, comme par exemple la structure du système de voisinage. Les avis d'experts peuvent permettre de s'affranchir de manière ad hoc de cette difficulté mais cela reste subjectif. Nous nous sommes intéressés au critère BIC pour la sélection de modèle pour les champs de Markov cachés. Si pour des modèles de mélanges indépendants ce critère est accessible directement, ce n'est plus le cas avec les modèles de champs de Markov et champs de Markov cachés. Le calcul de l'estimateur du maximum de vraisemblance (EMV) et de la vraisemblace des paramètres nécessite des approximations. Nous proposons, au vu des résultats de nos travaux sur les algorithmes EM de type champ moyen (*cf.* 6.1.3), d'utiliser l'algorithme en champ simulé pour approcher l'EMV. Puis, à partir d'une reformulation de BIC en termes de constantes de normalisation de la distribution des champs de Markov, nous proposons une approximation de ce critère basée sur l'approximation en champ moyen à l'ordre 1 de la constante de normalisation d'une distribution de Gibbs. Cela conduit à un critère simple et permet ainsi d'éviter la lourdeur des méthodes de type chaînes de Markov de Monte-Carlo (MCMC) [CFP00]. Ce critère montre de bonnes performances dans le cadre de la sélection du nombre de classes sur des données simulées et réelles.

[CFP00] G. Celeux, F. Forbes, N. Peyrard, « Mean field approximation principle for parameter estimation in hidden Markov models », *in : First European Conference on Spatial and Computational Statistics*, Ambleside, Grande Bretagne, 17-21 septembre 2000.

6.2.2 Combinaison de modèles en analyse discriminante

Participants : Isabel Brito, Gilles Celeux.

Travail en collaboration avec Ana Maria Sousa Ferreira (université de Lisbonne).

Ce thème constitue le sujet des thèses d'Isabel Brito, pour les méthodes de discrimination sur variables quantitatives, et d'Ana Maria Sousa Ferreira, qui a considéré des modèles de discrimination sur variables qualitatives [12]. Le but de la combinaison de méthodes de discrimination est l'obtention de règles de décison à la fois plus stables et aussi performantes que celles tirées d'une seule méthode. Cette année, Isabel Brito a essentiellement exploré la combinaison hiérarchique qui revient à traiter le problème en une suite de problèmes à deux classes, l'art consistant à choisir de manière pertinente les classes considérées à chaque niveau de la construction hiérarchique. De plus, cette technique a été appliquée à un problème de reconnaissance de tumeurs du cerveau. Dans le cadre qualitatif où travaille Ana Maria Sousa Ferreira, le couplage hiérarchique a aussi été employé et donne des résultats particulièrement intéressants.

6.2.3 Analyse discriminante sur tableaux de dissimilarités

Participants : Gilles Celeux, Anne Guérin-Dugué.

On considère la situation où chaque groupe à classer n'est pas connu par un ensemble de descripteurs mais par des indices de proximité ou de dissimilarité des groupes entre eux. Ce type de structure de données se rencontre couramment en psychophysique, biologie..., mais aussi en analyse d'image et du signal.

Nous avons continué notre recherche sur l'analyse discriminante sur tableaux de dissimilarités. Les premiers résultats ont permis de valider l'approche sur des bases de donnée réelles pour la classification de protéines [43] ou la catégorisation d'images naturelles [42]. Les algorithmes proposés ont été étendus pour un nombre de groupes supérieur à deux [61]. De plus pour rendre l'approche plus flexible, nous avons introduit la notion de coût d'erreur pour prendre en compte des probabilités a priori différentes pour chaque groupe. Un cas particulier assez courant, pour lequel les méthodes classiques sont en défaut, a été mis en exergue. Si pour un ensemble d'observations, les différents groupes de cette base appartiennent à des variétés de dimension intrinsèque différente, une analyse discriminante quadratique dans l'espace des caractéristiques conduira à des instabilités numériques. Par la méthode proposée, la prise en compte des "formes" et des dimensions intrinsèques différentes entre les groupes s'effectue implicitement par apprentissage de la métrique.

L'utilisation de cette technique dans les systèmes de recherche d'information est en cours d'étude.

6.3 Modèles de fiabilité industrielle

6.3.1 Un modèle de vieillissement

Participants : Henri Bertholon, Gilles Celeux.

Cette recherche a fait l'objet de la thèse de Henri Bertholon [11]. La thèse traite de l'existence et de la formalisation du vieillissement de systèmes que l'on rencontre dans le domaine de la fiabilité.
Comme entrée en matière nous avons étudié une approche bayésienne non informative du test de l'exponentialité contre une loi de Weibull. Nous proposons ensuite une nouvelle modélisation du vieillissement adaptée aussi bien aux cas des matériels réparables que non réparables. L'objectif de cette modélisation est double. D'un côté le modèle fait apparaître explicitement un paramètre correspondant à un instant de début du vieillissement, de l'autre il permet de dissocier les deux causes principales d'une défaillance, à savoir l'accident et le vieillissement. En définitive, la loi du modèle s'interprète, dans le cas d'un matériel non réparable, comme le minimum d'une loi exponentielle et d'une loi de Weibull décalée indépendantes. Dans le cas d'un matériel réparable, il s'agit de la superposition d'un processus de Poisson homogène et d'un processus de Weibull décalé indépendants. Dans ce cadre, nous avons développé une procédure d'estimation par le maximum de vraisemblance utilisant l'algorithme EM. Nous avons prouvé l'existence d'une solution convergente de l'équation de vraisemblance, dans le cas non réparable. Des simulations ont montré la cohérence de cette approche. En second lieu, nous avons proposé un test global de l'existence du vieillissement, pour lequel nous avons prouvé que la variable de décision a une loi indépendante du paramètre sous l'hypothèse d'absence de vieillissement. Nous construisons alors une table des seuils critiques pour différents risques α. Enfin nous examinons des applications réelles de notre modèle principalement dans le domaine industriel, mais aussi dans le domaine actuariel puisque nous proposons finalement l'analyse d'une table de mortalité qui met bien en évidence l'intérêt de notre modélisation.

6.3.2 Un modèle de choc

Participant : Gilles Celeux.

Cette recherche se fait en collaboration avec Andei Rodionov de l'IPSN. Il s'agit de caractériser la durée de vie d'un matériel soumis à des chocs lors de sollicitations et pouvant aussi être défaillant pour des raisons accidentelles indépendantes des chocs. Nous avons conçu un modèle à risques concurrents où les défaillances dues aux chocs obéissent à une loi Gamma. Il s'agit d'un modèle à risque masqué pour lequel nous avons développé beaucoup d'outils algorithmiques [1] dont nous tirons parti pour l'estimation des paramètres de ce modèle.

6.3.3 Modèle graphique et applications à la maintenance

Participants : Gilles Celeux, Franck Corset.

Dans le cadre d'une convention d'étude et de recherche avec le groupe « Sûreté, Diagnostic, Maintenance » de EDF-DER, nous nous intéressons à modéliser les dégradations d'un matériel

par un modèle graphique, appelé aussi réseau bayésien. L'idée consiste à proposer un modèle de graphe aléatoire acyclique prenant en compte de manière simple et explicite les facteurs fonctionnels pouvant influencer l'apparition de maladies sur des systèmes rentrant en jeu dans le fonctionnement des centrales nucléaires.

Après la construction de la structure du graphe, nous avons interrogé trois experts du département SDM afin d'évaluer les probabilités attachées aux arêtes ainsi constituées. Au vu du trop grand nombre de probabilités à fournir et ne disposant que de très peu de données de retour d'expériences, nous avons choisi de représenter le modèle graphique par un modèle log-linéaire[Whi90] non saturé, où les interactions d'ordre supérieur à 2 sont supposées nulles. De plus, nous avons demandé aux experts toutes les probabilités marginales et les probabilités conditionnelles des nœuds sachant un seul parent. Ainsi, le système d'équation à résoudre possédait plus d'équations que de variables. Ceci nous a permis de vérifier la cohérence des avis d'experts. En cas d'incohérence, des règles ont été proposées pour choisir les probabilités à garder. Puis nous avons réalisé une inférence pour chaque expert. Des scénarios critiques sont proposés en maximisant la probabilité que le composant soit défaillant par des algorithmes analogues à celui du MAP pour les chaînes de Markov.

6.3.4 Modélisation d'un changement de comportement de maintenance

Participants : Gilles Celeux, Franck Corset.

Suite à une convention d'étude et de recherche avec le groupe « Sûreté, Diagnostic, Maintenance » de EDF-DER, nous avons poursuivi l'étude suivante : il arrive que lors du suivi de la vie d'un système, les premières années les ingénieurs de maintenance mettent au rebut des matériels non sur une base objective mais par précaution excessive. De la sorte, les estimations des modèles de durée de vie sont entachées d'un biais pessimiste. Nous avons mis au point un modèle qui permet de supprimer ce biais d'estimation. Il consiste à voir là un problème à données cachées, l'information manquante étant de savoir si avant une date connue, les rebuts de matériels ont été faits par précaution. Ce modèle est estimé par le maximum de vraisemblance via l'algorithme EM ou par inférence bayésienne via l'échantillonnage de Gibbs. Nous avons observé que dans un cadre exponentiel, il est possible d'estimer les paramètres du modèle en maximisant directement la vraisemblance complète [14]. Les résultats expérimentaux sont satisfaisants et ont été confirmés par une validation par simulation de Monte-Carlo. Une perspective est de reprendre ce type de modèle pour des matériels soumis à un vieilissement régi par une loi de Weibull.

6.3.5 Modélisation et estimation de queues de distributions

Participants : Jean Diebolt, Myriam Garrido, Stéphane Girard.

Dans le cadre d'une convention d'étude et de recherche avec le groupe « Retour d'expérience » de EDF-DER, nous nous intéressons au problème de l'estimation des probabilités d'événements rares (ou de queues de distribution) et plus particulièrement à l'estimation de quantiles extrêmes — situés à proximité ou au-delà de la dernière observation ordonnée.

[Whi90] J. WHITTAKER, *Graphical Models in Applied Multivariate Statistics*, John Wiley, 1990.

Lors d'un premier contrat, il est ressorti que la méthode ET (*exponential tail*) pouvait être un moyen simple de réaliser l'estimation de ces quantiles. Un deuxième contrat a étudié le comportement asymptotique de cette méthode ; ce qui a notamment permis la mise en place d'un test d'adéquation de modèles paramétriques à la queue de distribution. En pratique, il arrive que les tests d'adéquation usuels (dépendant principalement de la partie centrale de la distribution) aboutissent à des conclusions différentes de celles de ce test extrême. Ainsi, un troisième contrat a proposé des procédures de régularisation de la loi obtenue de sorte qu'elle ne perde pas trop de son ajustement central mais qu'elle s'adapte mieux en queue de distribution. La méthode finalement retenue est celle de la régularisation bayésienne qui permet la prise en compte d'un avis d'expert. Le contrat[DDG00] a permis de continuer l'étude du test d'adéquation à la queue de distribution précédemment proposé. Nous avons étudié sa puissance du double point de vue des simulations et de la théorie. Nous avons aussi étendu la procédure de régularisation bayésienne au cas de la loi de Weibull avec une loi a priori sur le paramètre de forme. Ce cas est important, car c'est en changeant le paramètre de forme que l'on obtient les plus grandes modifications de la loi régularisée, mais difficile car il n'existe pas de loi conjuguée permettant un calcul analytique des lois a posteriori et prédictive, ce qui oblige à des calculs au cas par cas.

Actuellement, nous explorons dans le cadre du présent contrat [50, 55] une estimation bayésienne des lois de Pareto généralisées, lois limites de la loi des excès quand le seuil tend vers l'infini, qui permettent l'estimation des queues de distribution et des quantiles extrêmes. L'introduction d'une méthode bayésienne (notamment avec un avis d'expert sur les queues de distribution) pourrait nous permettre de réduire le biais inhérent à la méthode d'estimation des quantiles extrêmes. La loi prédictive a posteriori sera proposée pour estimer la queue de distribution. Nous étudions en relation avec EDF la manière de prendre au mieux en compte un avis d'expert dans ce contexte.

6.3.6 Application des chaînes de Markov cachées à la fiabilité de logiciels

Participant : Jean-Baptiste Durand.

Travail en collaboration avec Olivier Gaudoin et Jean-Louis Soler du LMC, Grenoble.

Cette étude porte sur la modélisation des temps inter-défaillance d'un logiciel. Le modèle est basé sur les hypothèses suivantes : le logiciel est sans usure ; après chaque défaillance, le logiciel subit éventuellement une correction susceptible de modifier son taux de défaillance ; les corrections apportées ne dépendent que de l'état actuel du logiciel ; le nombre de "versions" du logiciel est fini.

Ces hypothèses nous conduisent à considérer que les durées inter-défaillance obéissent à un modèle de Markov caché à lois conditionnelles exponentielles. Des techniques de choix de modèles basées sur des critères de type vraisemblance pénalisée ou validation croisée permettent de déterminer le nombre de versions significatives du logiciel (nombre d'états cachés) ainsi que le type de dynamique dans l'amélioration ou la dégradation du logiciel (matrice de transition). Le modèle est mis en compétition avec des modèles classiques de croissance de fiabilité, sur

[DDG00] J. Diebolt, V. Durbec, M. Garrido, « Extremes : logiciel d'analyse des événements extrêmes », rapport final de convention de recherche Inria-EDF, 2000.

un critère de capacité prédictive (graphes *uplot*). Ce critère permet de conclure que le modèle est compétitif par rapport aux autres dans le cas de logiciels à faible croissance de fiabilité ou dans le cas où la fiabilité est susceptible de décroître.

6.4 Statistique biomédicale

Participants : Christine Cans, Gilles Celeux, Cécile Delhumeau, Olivier Martin, Christian Lavergne, Claudine Robert.

6.4.1 Analyse de données issues de puces à ADN

Participants : Gilles Celeux, Olivier Martin.

Les puces à ADN [FOR99] permettent de mesurer simultanément le niveau d'expression de plusieurs milliers de gènes. Le volume et la complexité des données obtenues rendent les études difficiles et nécessitent des développements statistiques. Les domaines que nous privilégions concernent les problèmes de normalisation et de recherche de gènes différentiellement exprimés. En collaboration avec l'IPMC (Institut de Pharmacologie Moléculaire et Cellulaire) à Nice, nous avons implémenté les méthodes développées par Yang *et al.* [YDLS01] permettant de supprimer certains effets liés à la technologie et la biologie. Les premiers résultats ont montré la nécessité de définir un plan d'expérience. La recherche de gènes différentiellement exprimés [NKR+01] permet d'identifier les gènes qui sont impliqués dans un processus biologique. Sur ce point, nous avons proposé [44] une approche bayésienne qui modélise l'induction, la répression et la non différenciation. Ce modèle a été utilisé afin d'analyser les résultats de l'IMPC sur des kératinocytes humains et sur des kératinocytes canins pendant une phase de réparation cellulaire.

6.4.2 Évolution d'indicateurs périnatals

Participants : Christine Cans, Cécile Delhumeau, Christian Lavergne.

La « Cerebral Palsy » (CP) ou infirmité motrice d'origine cérébrale est une maladie infantile qui compromet l'autonomie de l'enfant et induit une prise en charge lourde. C'est l'une des maladies les plus commune chez les jeunes enfants. Un projet européen a pour but de fournir des estimations fiables du taux de prévalence de la CP en Europe et d'identifier les facteurs de risque de cette maladie. Dans ce cadre, nous comparons des régions d'Europe selon des indicateurs périnatals observés dans la population générale : taux de faible poids ($< 1\,500$g) de naissance unique, taux de naissances multiples, taux de morti-natalité, et taux de décès néonatal chez les enfants pesant moins de 1 500g. Les résultat des analyses montrent que

[FOR99] *The Chipping Forecast*, *21*, 1999. Supplement to Nature Genetics.

[YDLS01] Y. Yang, S. Dudoit, P. Luu, T. Speed, « Normalization for cDNA microarray data », *rapport de recherche*, Statistics Dept, UC Berkeley, 2001.

[NKR+01] M. Newton, C. Kendziorski, C. Richmond, F. Blattner, K. Tsui, « On differential variability of expression ratios: Improving staistical inference about gene expression changes from microarray data », *Journal of Computational Biology 8*, 2001, p. 37–52.

l'évolution des indicateurs périnataux étudiés est propre à chaque pays, à chaque année et quelquefois à chaque sexe car elle est très liée aux techniques médicales utilisées.

6.4.3 Analyse des durées de séjour du CHU de Grenoble

Participants : Gilles Celeux, Cécile Delhumeau.

Travail en collaboration avec Jérôme Fauconnier (CHU Grenoble).

Dans le cadre du Programme Médicalisé des Systèmes d'Informations (PMSI) qui sert à évaluer l'activité des hôpitaux et à ajuster leurs budgets, chaque établissement produit pour chaque séjour d'un patient un résumé standardisé de sortie, qui résume les principales données médico-sociales de son séjour. À partir de ces données, les séjours sont regroupés en Groupes Homogènes de Malades (GHM), qui sont en quelques sorte l'unité de production hospitalière. Notre étude vise à comparer les distributions des durées de séjour (DS) des GHM du CHU de Grenoble à celles de la bases de données nationale (sondage à 5%), afin de mettre en évidence d'éventuels dysfonctionnements au sein d'un service, des recrutements ou des prises en charges différents de patients à Grenoble où les durées de séjour ont tendance à être plus longues. Après une détection des GHM grenoblois singuliers par leurs DS, à l'aide notamment du logiciel MIXMOD, nous travaillons actuellement l'analyse des caractéristiques médico-sociales des patients des différents groupes de GHM obtenus.

6.5 Inférence statistique pour le traitement du signal et des images

6.5.1 Analyse d'images

Participants : Anne Guérin-Dugué, Paulo Gonçalvès.

Travail en collaboration avec Paulo Oliveira et Victor Barroso de l'ISR-IST, Lisbonne.

Durant cette première année, nous avons fait une recherche bibliographique concernant l'extension de la notion de signal analytique en 2 dimensions. Une voie intéressante a été identifiée. Elle concerne une extension dans l'espace des quaternions [Bul99]. Le point clé des extensions proposées est le traitement conjoint des 2 dimensions spatiales, ou fréquentielles dans le domaine de Fourier, de manière combinée. La propriété principale qui en découle est l'isotropie. Des programmes de simulation ont été établis et testés sur des images artificielles.

La suite de cette recherche va consister à mutualiser différents travaux pour former une chaîne de traitement complète :

1. Décomposition des images par *Empirical Mode Decomposition* (EMD) [H+98], (P. Gonçalvès, P. Oliveira). Une extension 2D de l'algorithme original a déjà été développée dans le cadre de cette collaboration.
2. Analyse locale des éléments de la décomposition (A. Guérin-Dugué, P. Oliveira)

[Bul99] T. Bulow, *Hypercomplex Spectral Signal Representations for the Processing and Analysis of Images*, thèse de doctorat, Ph.D. from *Christian-Albrechts-Universität*, Kiel, 1999.

[H+98] N. E. Huang *et al.*, « The Empirical Mode Decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis », *The Royal Society*, 1998.

6.5.2 Synthèse de processus multifractals

Participant : Paulo Gonçalvès.

Travail en collaboration avec Rudolf Riedi (Rice university, Houston (TX), USA).

Nous avions conduit dans [GR99] l'étude des statistiques au second ordre de la décomposition en ondelettes des mouvements browniens fractionnaires en temps multifractal. Ces processus multifractals (i.e. possédant des spectres de singularité non dégénérés), sont pressentis comme étant des bons modèles pour caractériser et simuler une large gamme de signaux réels : séries financières, charges de réseaux informatiques, signaux biologiques et physiologiques (rythme cardiaque notamment)...

Nous proposons à présent un méthode de *synthèse exacte* de trajectoires de ces processus, valable pour toute valeur du paramètre de longue dépendance (encore appelé paramètre de Hurst) compris dans l'intervalle $[0, 1]$. Cette procédure simple et rapide permet de générer des traces de grande taille (2^{16} points en moins de 2 secondes). Schématiquement, l'algorithme se décompose en quatre étapes :

1. synthèse d'une trajectoire de mouvement brownien fractionnaire de paramètre $0 < H < 1$, sur N points, par la méthode de la matrice circulante [CB99] ;
2. décomposition en ondelettes orthogonales de cette trace ;
3. pondération des coefficients d'ondelette par les multiplicateurs de la mesure multinomiale "temps" [GR99] ;
4. synthèse de la trajectoire du mouvement brownien fractionnaire en temps multifractal par inversion de la décomposition en ondelette.

6.5.3 Test d'existence des moments d'ordre q d'une variable aléatoire

Participant : Paulo Gonçalvès.

Travail en collaboration avec Rudolf Riedi (Rice university, Houston (TX), USA) et Anestis Antoniadis (IMAG-LMC).

Pour un mouvement brownien fractionnaire (processus gaussien), on sait que les coefficients d'ondelette $C_{n,k}$ (eux-même gaussiens) élevés à la puissance q, avec $q < -1$, suivent une loi $\gamma-$stable (d'espérance infinie), et l'estimateur empirique de moments $S^n(q)$ ne converge donc pas vers une limite finie. Dans ce cas, un test de stabilité décide d'affecter la valeur $+\infty$ à la variable $S^n(q)$ [GRB98]. En pratique, lorsque l'on ne connait pas la loi de distribution des coefficients d'ondelette, on ne sait pas a priori pour quels ordres de q les moments de la v.a $|C_{n,k}|$ existent.

[GR99] P. Gonçalvès, R. Riedi, « Wavelet Analysis of Fractional Brownian Motion in Multifractal Time », *in : Proceedings of the 17th Colloquium GRETSI*, Vannes, France, septembre 1999.

[CB99] M. S. Crouse, R. G. Baraniuk, « Fast, Exact Synthesis of Gaussian and nonGaussian Long-Range-Dependent Processes », *IEEE Transactions on Information Theory*, 1999, Submitted.

[GRB98] P. Gonçalvès, R. Riedi, R. Baraniuk, « A Simple Statistical Analysis of Wavelet-based Multifractal Spectrum Estimation », *in : Proceedings of the 32nd Conference on "Signals, Systems and Computers"*, Asilomar, USA, Nov. 1998.

Nous avions proposé dans [Gon00], un test simple permettant de déterminer le domaine d'existence (q_{min}, q_{max}) de la fonction $S^n(q)$. Le test proposé était basé sur l'analyse par ondelettes de la régularité ponctuelle de la fonction caractéristique de la variable aléatoire.

Nous avons depuis approfondi cette étude, en déterminant le biais asymptotique de l'estimateur, ainsi que sa variance. Ce faisant, nous mettons en lumière le rôle du choix de l'ondelette d'analyse (notamment à travers sa régularité), ainsi que la confiance à accorder aux valeurs extrêmes, compte tenu de la taille N de l'échantillon traité.

Avec A. Antoniadis, nous commençons une étude plus spécifique de ce test d'existence de moments, appliqué à des variables α-stables, pour lesquelles des estimateurs fiables existent déjà (Koutrouvelis, Mac Cullogh,...). D'une certaine façon, on peut envisager les *estimateurs ondelettes* proposés comme la généralisation de ces méthodes à des estimateurs à noyaux de fonctions caractéristiques.

6.5.4 Diffusion de représentations temps-fréquence pour un problème décisionnel

Participant : Paulo Gonçalvès.

Travail en collaboration avec Julien Gosme (Université de Technologie de Troyes).

Nous avons développé dans le cadre d'un stage de DEA, une nouvelle méthode permettant de contrôler de façon souple et efficace, la complexité des détecteurs temps-fréquence (TF). Le schéma proposé, inscrit dans le contexte des détecteurs construits sur des bases d'apprentissage, repose sur le concept récemment introduit de diffusion non uniforme des représentations TF [GP98]. Inspirée de l'analyse multi-échelle utilisée en traitement d'images [LDAR97], cette diffusion a été transposée à l'analyse temps-fréquence, pour permettre d'effectuer un lissage localement adaptatif (en temps et en fréquence) du signal. Dans ce travail, nous avons exploité cette propriété d'adaptativité locale, pour définir une classe de détecteurs TF non paramétriques qui augmentent singulièrement le contraste entre les hypothèses en compétition. Pour contrôler le processus itératif de diffusion, nous utilisons par ailleurs un critère d'arrêt directement lié à la probabilité d'erreur du détecteur. Cela nous permet de retenir dans la classe de ces détecteurs TF, celui produisant la meilleure décision [56].

6.6 Commande adaptative

Participant : Anatoli Iouditski.

Travail en collaboration avec Alexandre Nazin (IPU, Moscou, Russie).

Par rapport aux problèmes classiques d'estimation stochastique, le problème de la commande adaptative est assez singulier : il possède un degré de liberté supplémentaire qui est

[Gon00] P. Gonçalvès, « Existence test of moments: Application to Multifractal Analysis », *in: Proceedings of Int. Conf. on Telecom.*, mai 2000.

[GP98] P. Gonçaalvès, E. Payot, « Diffusion equation for time frequency representation », *Proc. IEEE Digital signal processing workshop*, 1998.

[LDAR97] L. Lucido, R. Deriche, L. Alvarez, V. Rigaud, « Sur quelques schémas numériques de résolution d'equations aux dérivées partielles pour le traitement d'images », *rapport de recherche n° RR-3192*, Inria, Institut National de Recherche en Informatique et en Automatique, 1997.

la commande. Nous avons continué, en collaboration avec des chercheurs de l'IPU (Institute for Control Science) de Moscou, l'étude des algorithmes adaptatifs de commande pour des systèmes dynamiques non linéaires. Des nouveaux algorithmes de commande ont été proposés et leur efficacité a été établie (cf. [27]).

6.7 Estimation de paramètres macroscopiques

Participant : Anatoli Iouditski.

Travail en collaboration avec Marian Hristach (ENSEI, Rennes), et Vladimir Spokoiny (WIAS, Berlin, Allemagne).

Dans un bon nombre de problèmes de modélisation non paramétrique appliqués à la finance la question qui intéresse le chercheur est d'établir une estimation de paramètres macroscopiques ou spatiaux de modèles inconnus. Un exemple classique est le problème d'estimation de l'indice spatial dans le modèle de type "single-indice" ou multi-indice. Nous proposons [22] une nouvelle méthode d'estimation du coefficient d'indice dans un modèle "single-indice", qui est basée sur des améliorations itératives de l'estimateur de la dérivée moyenne. L'estimée qui en résulte est $\sqrt{n}$–consistente, n étant la taille de l'échantillon. Dans [21] nous avons généralisé cette méthode d'estimation très prometteuse dans le cadre de modèles multi-indices.

6.8 Intervalles de confiance pour des algorithmes adaptatifs

Participant : Anatoli Iouditski.

Travail en collaboration avec Oleg Lepski (Université Aix-Marseille I), et Sophie Lambert-Lacroix (Université Joseph Fourier).

Nous nous sommes intéressés à la construction d'intervalles de confiance pour des estimateur fonctionnels adaptatifs. Nous avons établi le cadre théorique précis d'estimation de la norme d'erreur d'estimation et d'intervalle de confiance associés. En outre, nous avons proposé des algorithmes adaptatifs d'estimation non paramétrique par ondelettes avec des intervalles de confiance associés. On démontre l'efficacité de ces algorithmes sur une grande variété de classes fonctionnelles [23, 24].

7 Contrats industriels (nationaux, européens et internationaux)

7.1 Utilisation de modèles graphiques en fiabilité

Participants : Gilles Celeux, Franck Corset.

Ce contrat [53] de type CRECO avec le département « Surveillance, Diagnostic, Maintenance » de la DRD-EDF Chatou concernait l'utilisation de modèles graphiques en fiabilité et pour l'optimisation de maintenance. Cette étude faisait suite à une étude de même type effectuée l'an dernier. Cette année nous nous sommes concentrés sur les possibilités d'entrer les probabilités nécessaires à l'inférence dans un modèle graphique lorsque le nombre de données est très faible et que l'on doit recourir aux avis d'experts. La deuxième partie du contrat consiste à

donner les scénarios critiques et les variables susceptibles de jouer un rôle discriminant. Enfin, un logiciel a été réalisé grâce à la boîte à outil BNT de Matlab.

7.2 Contrat EDF sur les queues de distribution de probabilité

Participants : Jean Diebolt, Myriam Garrido, Stéphane Girard.

Ce contrat de type GRECO entre IS2 et le groupe « Retour d'expérience » de EDF-DER porte sur l'estimation des queues de distributions et des quantiles extrêmes au-delà de la plus grande valeur d'un échantillon. Plus précisément, si X est une variable aléatoire, le problème peut se résumer à l'estimation du quantile q_{1-p_n} défini par :

$$P(X > q_{1-p_n}) = p_n, \;\; p_n \leq 1/n.$$

Cette étude prolonge le travail des trois années précédentes sur ce même thème. Nous disposons maintenant de tests permettant de vérifier l'adéquation d'un modèle paramétrique à un échantillon, tant du point de vue de sa forme globale que du point de vue de sa queue de distribution. Nous avons aussi complété la procédure de régularisation bayésienne proposée l'an dernier dans le cadre d'un autre contrat. Nous explorons la piste bayésienne pour l'estimation des lois de Pareto generalisées (voir 6.3.5). Une Journée EDF de formation au logiciel Extremes est envisagée pour fin 2001 ou début 2002.

7.3 Étude de courbes de consommation électrique

Participants : Gilles Celeux, Jean-Baptiste Durand.

Ce contrat de type CRECO avec le département «Clientèle» de EDF-DRD CLAMART a pour objet l'analyse statistique de courbes de consommation électrique. L'hypothèse qui préside à la modélisation est que la consommation d'un ménage dépend essentiellement d'un état non observé lié au type d'activité (repas, veillée, sommeil, etc.) et que ces états obéissent à un régime markovien.

La première étape de l'analyse a consisté à estimer et à isoler tous les effets non aléatoires tels que la saisonnalité et les effets dus à l'heure de la journée ou au type de contrat, par une analyse de la variance sur les log-consommations. Cette partie a fait l'objet du stage de Mathieu Thivin de l'Enserg. Les résidus de cette analyse de variance sont alors modélisés par une chaînes de Markov cachés. Cette phase du travail est en cours.

7.4 Scénarios de défaillance de pénétration de fonds de cuves

Participants : Gilles Celeux, Guillaume Bouchard.

Ce contrat de type CRECO avec le groupe «Fiabilité des composants et structures» de EDF-DRD CHATOU a pour objet la mise en place d'une modélisation bayésienne pour l'élaboration de scénarios d'amorçage et de propagation de fuites sur les fonds de cuves REP. Ces scénarios doivent prendre en compte les modèles mécaniques probabilistes de rupture et les connaissances a priori. Cela conduit à des modèles compliqués, qui dans une optique bayesienne, doivent être identifiés à l'aide d'algorithmes MCMC.

7.5 Contrat CEA (Cadarache) : Étude d'incertitudes et de sensibilité

Participants : Christian Lavergne, Cyril Goutte, Gérard Boudjema, Julien Jacques.

Nous avons continué la collaboration avec le CEA/DER, dans laquelle nous apportons notre expérience sur le développement de méthodes pemettant de maîtriser les incertitudes et de déterminer les paramètres les plus influents dans des processus complexes. Ces travaux sont menés dans le cadre de deux applications : Le Programme de Suivi des Irradiations (PSI) des cuves de réacteurs et les scénarios d'accidents graves.

En 2001, deuxième année de la collaboration, nous avons développé des outils permettant d'évaluer la sensibilité des paramètres d'entrée sur un système non linéaire et non monotone : le code Actige. Deux études de sensibilité ont été effectuées : une concernant le calcul des incertitudes par le code Circé, une autre sur la décomposition de Sobol pour des distributions non uniformes des entrées. D'autre part une analyse de code STAY'SL a été démarrée et se fait en deux étapes : la première qui est actuellement achevée a porté sur le processus de préparation des données, la deuxième portera sur le code de calcul proprement dit.

Cette activité a été menée par Cyril Goutte (du 01/01 au 30/09) et est poursuivie par Gérard Boudjema depuis le 01/10 et ce jusqu'à fin 2002. La partie "analyse de sensibilité" se poursuit par la thèse de Julien Jacques cofinancée par le CEA à compter du 1er novembre 2001.

8 Actions régionales, nationales et internationales

8.1 Actions régionales

IS2 participe régulièrement au séminaire de statistique du LMC-SMS à Grenoble et G. Celeux est l'un des organisateurs. Dans ce cadre, plusieurs conférenciers ont été invités. De plus, cette année, J.-B. Durand, N. Peyrard, et A. Iouditski ont exposé à ce séminaire.

G. Celeux est le représentant pour Rhône-Alpes du thème « Analyse de données d'expression » du comité bio-informatique des génopoles.

P. Gonçalvès participe à deux projets du programme de thématiques prioritaires de la région Rhône-Alpes. L'un, intitulé « Application de l'Analyse en Ondelettes à l'Acoustique et à la Turbulence » est placé sous la responsabilité de V. Perrier, Professeur à l'Ensimag (INPG), l'autre intitulé « Diagnostic Acoustique de la Vorticité dans les Écoulements Turbulents » sous la responsabilité de C. Baudet, Professeur à l'UJF (Legi).

Le groupe FIMA qui a fait l'objet d'un soutien par une ARC local a été créé. Ce groupe a pour but de fédérer ces activités de recherche entre le LMC et l'INRIA pour d'une part renforcer la visibilité du pôle grenoblois de recherche en fiabilité, et d'autre part développer de nouveaux axes de recherche. En particulier, nous souhaitons développer les relations entre les deux organismes participants et tous les partenaires locaux intéressés par la sûreté de fonctionnement, aussi bien les laboratoires de recherche que les entreprises. La principale activité de FIMA est le groupe de travail qui se réunit approximativement une fois par mois et auquel les membres d'IS2 impliqués en fiabilité participent activement.

8.2 Actions nationales

P. Gonçalvès entretient une collaboration régulière (2 jours par mois) avec l'équipe U127 de l'Inserm à l'hôpital Lariboisière (Paris), sur l'analyse du rythme cardiaque.

Depuis cette année, une collaboration a été initiée avec Yann Guédon du Cirad. Elle va se concrétiser par deux thèses qui viennent de démarrer. Elle concerne d'une part, des recherches en modèles graphiques à structures cachées. Ce thème fait l'objet de la thèse de Guillaumne Bouchard qui sera co-encadrée par Yann Guédon et Gilles Celeux. Elle concerne d'autre part, des recherches sur l'analyse de données longitudinales dans le cadre de la structure exponentielle. Ce thème fait l'objet de la thèse de Carine Véra (ASC INRA), co-encadrée par Yann Guédon et Christian Lavergne.

8.3 Réseaux et groupes de travail internationaux

G. Celeux, J. Diebolt et F. Forbes participent au réseau européen *Spatial and computational statistics*. Ils sont rattachés au nœud de Rouen animé par Ch. Robert (Crest).

8.4 Relations bilatérales internationales

Europe

G. Celeux poursuit sa collaboration avec le LEAD de l'université de Lisbonne, qui s'est traduite cette année par la soutenance de la thèse de Ana Maria Ferreira [12].

P. Gonçalvès, G. Celeux et A. Guérin-Dugué ont obtenu un financement de la part du programme de coopération scientifique bilatérale INRIA / ICCTI (Portugal). L'Institut des Systèmes et Robotique de l'Institut Supérieur de Technologie (Lisbonne) est notre collaborateur au Portugal, et avec lui nous menons une étude sur le thème *Inférences statistiques en Traitement du Signal : Mesures spectrales instantanées et modèles de mélanges gaussiens.*

Maghreb

G. Celeux poursuit des relations de recherche régulières avec A. Mkhadri (université de Marrakech). A. Mkhadri a passé deux mois à l'Inria Montbonnot et a participé à des recherches sur la sélection de modèles par validation croisée avec G. Celeux.

Amérique du Nord

Le projet IS2 poursuit sa collaboration avec le département de statistique de l'université de Washington à Seattle. F. Forbes a effectué un séjour de deux mois dans ce département et a exposé dans le groupe de travail « Model-Based Clustering and Applications » organisé par A. Raftery.

P. Gonçalvès travaille avec R. Riedi de l'Université de Rice (Houston, TX) sur la synthèse de processus multifractals et l'élaboration de tests d'existence des moments d'ordres supérieurs pour des variables aléatoires leptokurtiques (travail commun avec A. Antoniadis du LMC-IMAG).

9 Diffusion de résultats

9.1 Animation de la communauté scientifique

G. Celeux a organisé la Journée Didactique de mars du projet IS2 sur les mélanges de lois de probabilité (`http://www.inrialpes.fr/is2`).

P. Gonçalvès est responsable avec C. Doncarli (IrCyn, ECN) du groupe de travail *Analyse et Décision en Signal* du GDR-PRC ISIS (CNRS). P. Gonçalvès est également animateur avec P. Abry (ENS-Lyon) de l'Opération Thématique *Ondelettes et Fractales pour le Traitement du Signal et des Images* dans le cadre de ce même GDR-PRC ISIS.

P. Gonçalvès est coordonnateur avec P. Abry (ENS-Lyon) et J. Lévy-Véhel (projet Fractales de l'UR de Rocquencourt) d'un volume *Lois d'échelle, Fractales et Ondelettes* dans la collection *Information-Commande-Communication* éditée par Hermès Science Publications (Paris).

9.2 Enseignement universitaire

G. Celeux enseigne les méthodes d'analyse statistique multidimensionnelle dans le DEA d'instrumentation biologique et médicale de Grenoble.

J. Diebolt assure un cours au DEA de mathématiques appliquées à l'université de Marne-la-Vallée. Ce cours porte sur la fiabilité et les valeurs extrêmes.

P. Gonçalvès assure un cours de 17h30 sur *Temps-fréquence et analyse multirésolutions* en troisième année de l'ENSERG.

De plus, tous les membres du projet donnent des cours de statistique dans différentes filières de premier et de deuxième cycles.

9.3 Participation à des colloques, séminaires, invitations

G. Celeux a été conférencier invité au *workshop in statistical mixtures and latent structure modelling* d'Edinbourg, à la conférence ASMDA 2001 à Compiègne et à la conférence *Recent Developments in Mixture Models* d'Hambourg.

G. Celeux, J.-B. Durand, M. Garrido, S. Girard, A. Guérin-Dugué, et O. Martin ont participé aux XXXIIIèmes journées de statistique de la SfdS, à Nantes en mai 2001 où G. Celeux a organisé une session à la mémoire d'André Carlier.

I. Brito a participé à JOCLA 2001, rencontre de la société portugaise de classification, à Porto en mars.

N. Peyrard a présenté ses travaux dans le cadre des séminaires ENSAM/INRA/UMII à Montpellier, en octobre.

G. Celeux, F. Forbes, P. Gonçalvès, A. Ioudisky et C. Lavergne ont participé aux Journées de Statistiques de l'Inria en novembre.

J. Diebolt, M. Garrido et S. Girard ont participé à EXTREMES 2001 à Leuven (Belgique) en août.

10 Bibliographie

Ouvrages et articles de référence de l'équipe

[1] M. Bacha, G. Celeux, E. Idée, A. Lannoy, D. Vasseur, *Estimation de modèles de durées de vie fortement censurés*, Eyrolles, Paris, 1998.

[2] C. Biernacki, G. Celeux, G. Govaert, « Assessing a mixture model for clustering with the integrated completed likelihood », *IEEE Trans. on PAMI*, 2000, p. 267–272.

[3] G. Celeux, J. Diebolt, « A stochastic approximation type EM algorithm for the mixture problem », *Stochastic and Stochastics Reports 41*, 1992, p. 119–134.

[4] G. Celeux, G. Govaert, « Gaussian parsimonious clustering models », *Pattern Recognition 28*, 1995, p. 781–793.

[5] M. L. Comer, E. J. Delp, « The EM/MPM Algorithm for Segmentation of Textures Images : Analysis and Further Experimental Results », *IEEE Transactions on Image Processing 9*, 10, 2000, p. 1731–1744.

[6] F. Forbes, A. E. Raftery, « Bayesian Morphology : Fast Unsupervised Bayesian Image analysis », *Journal of the American Statistical Association 94*, 446, June 1999, p. 555–568.

[7] O. François, C. Lavergne, « Design for Evolutionary Algorithms - A Statistical Perspective », *IEEE Transactions on Evolutionary Computation 5*, 2001, p. 129–148.

[8] P. Gonçalvès, R. Riedi, « Wavelet Analysis of Fractional Brownian Motion in Multifractal Time », *in : Proceedings of the 17th Colloquium GRETSI*, Vannes, France, septembre 1999.

[9] A. Juditsky, H. Hjalmärsson, A. Benveniste, B. Delyon, L. Ljung, J. Sjöberg, Q. Zhang, « Non-linear black-box modelling in system identification : mathematical foundations », *Automatica 31(12)*, 1995, p. 1725–1750.

[10] C. Robert, *Méthodes statistiques pour l'I.A. ; l'exemple du diagnostic médical*, Masson, Paris, 1991.

Thèses et habilitations à diriger des recherches

[11] H. Bertholon, *Un modèle de vieillissement*, thèse de doctorat, Université Joseph Fourier, 2001.

[12] A. M. Ferreira, *Combinaçao de modelos em analise discriminante sobre variaveis qualitativas*, thèse de doctorat, Université nouvelle de Lisbonne, 2001.

[13] N. Peyrard, *Approximations de type champ moyen des modèles de champ de Markov pour la segmentation de données spatiales*, thèse de doctorat, Université Joseph Fourier, 2001.

Articles et chapitres de livre

[14] G. Celeux, F. Corset, M.-G. Garnero, C. Breuils, « Accounting for inspection errors and change in maintenance behaviour », *Journal of Management Mathematics Special Issue*, 2001, À paraître.

[15] G. Celeux, F. Forbes, A. Mkhadri, S. Chrétien, « A Component-Wise EM algorithm for Mixtures », *Journal of Computational and Graphical Statistics*, December 2001, À paraître.

[16] G. Celeux, F. Forbes, N. Peyrard, « EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation », *To appear in Pattern Recognition*, 2002.

[17] G. Celeux, « Situations de maintenance à structure de données incomplètes », *Journal de la SFdS 141*, 3, 2001, p. 43–59, À paraître.

[18] J. Diebolt, M. El Aroui, « On the use of the peaks over threhold method for estimating out-of-sample quantiles », *Computational Statistics and Data Analysis*, 2001, à paraître.

[19] J. Diebolt, J. Zuber, « On testing the goodness-of-fit of a nonlinear heteroscedastic regression models », *Communications in Statistics – Simulation and Computation*, 2001.

[20] P. Flandrin, P. Gonçalvès, P. Abry, « Analyses en ondelettes et lois d'échelle », *in : Lois d'échelle, Fractales et Ondelettes*, *Information-Commande-Communication*, Hermes Science, 2001, à paraître.

[21] M. Hristache, A. Juditsky, A. Polzehl, V. Spokoiny, « Structure adaptive approach for dimension reduction », *Ann. of Stats*, 2001.

[22] M. Hristache, A. Juditsky, V. Spokoiny, « Direct Estimation of the Index Coefficients in a Single-index Model », *Ann. of Stats*, 2001.

[23] A. Juditsky, O. Lambert-Lacroix, « On nonparametric confidence set estimation », *Math. Meth. of Stat*, 2002, à paraître.

[24] A. Juditsky, O. Lepski, « Evaluation of the accuracy of nonparametric estimators », *Math. Meth. of Stat*, 2001, à paraître.

[25] A. Juditsky, A. Nemirovski, « On Nonparametric Tests of Positivity / Monotonicity / Convexity », *Ann. of Stats*, 2002, à paraître.

[26] J. Labarere, P. François, P. Auquier, C. Robert, M. Fourny, « Development of a French inpatient satisfaction questionnaire », *International Journal for quality of health care 13*, 2, 2001, p. 99–108.

[27] A. Nazin, A. Juditsky, « On minimax approach to nonparametric adaptive control », *Int. J. of Adaptive Contr. and Signal Proc.*, 2001.

[28] J.-P. Ovarlez, P. Gonçalvès, R. Baraniuk, « Analyse temps-fréquence quadratique III : La classe affine et autres classes covariantes », *in : Temps-fréquence : concepts et outils*, *Information-Commande-Communication*, Hermes Science, 2001, à paraître.

[29] Y. Vernaz, « Normalité asymptotique de l'estimateur des moindres carrés généralisées dans les modèles ARCH », *Note au CRAS - Serie I/Mathématique*, 2001, À paraître.

Communications à des congrès, colloques, etc.

[30] I. Brito, G. Celeux, « Combination in Gaussian discriminant analysis : Some notes », *in : JOCLAD2001*, Porto, février 2001.

[31] C. Cans, C. Delhumeau, C. Lavergne, « Étude de la mortalité néonatale, de 1980 à 1990, chez les enfants de moins de 2500g à la naissance, dans 3 pays européens. », *in : XXXIèmes journées de la société française de médecine périnatale*, Lille, 25-26 octobre 2001.

[32] G. Celeux, « Assessing the number of mixture components : a survey », *in : Workshop in statistal mixtures and latent structure modelling*, 2001.

[33] G. Celeux, « Different points of view for choosing the number of components in a mixture model », *in : Applied Stochastic Modeles and Data Analysis*, G. Govaert, J. Janssen, N. Limios (éditeurs), p. 21–29, Compiègne, juin 2001.

[34] G. Celeux, « mixmod : a Software for Gaussian Mixture Estimation », *in : Recent Developments in Mixture Modelling*, 2001.

[35] J. Diebolt, « A new Bayesian approach to GPD's », *in : Second International Symposium on Extreme Value Analysis*, Leuven, Belgique, août 2001.

[36] J. B. Durand, « Choisir l'ordre d'une chaîne de Markov cachée par Half Sampling. », *in : XXXIIIèmes journées de la SFdS*, Nantes, France, 14-18 mai 2001.

[37] A. Ferreira, G. Celeux, H. Bacelar, « A Hybrid model in discrete discriminant analysis », *in : JOCLAD2001*, Porto, février 2001.

[38] M. Garrido, J. Diebolt, S. Girard, « The ET test, a goodness-of-fit test to the distribution tail », *in : Second International Symposium on Extreme Value Analysis*, Leuven, Belgique, août 2001.

[39] M. Garrido, J. Diebolt, « A Bayesian regularisation procedure for a better extremal fit », *in : Second International Symposium on Extreme Value Analysis*, Leuven, Belgique, août 2001.

[40] S. Girard, P. Jacob, « Extreme values estimates of point processes boundaries », *in : Second International Symposium on Extreme Value Analysis*, Leuven, Belgique, juillet 2001.

[41] A. Guérin-Dugué, C. Biernacki, J. Hérault, « Statistical modelling for image retrieval using a biological model of the perceptive colour space », *in : ICIP*, Thessaloniki, Greece, october 2001.

[42] A. Guérin-Dugué, G. Celeux, « Analyse discriminante sur tableau incomplet de dissimilarités », *in : XXXIIIèmes journées de statistique*, Nantes, 2001.

[43] A. Guérin-Dugué, G. Celeux, « Discriminant Analysis on Dissimilarity Data : A New Fast Gaussian-like Algorithm », *in : Artificial Intelligence and Statistics 2001*, 2001.

[44] O. Martin, « Puce à ADN et analyse de l'expression des gènes », *in : XXXIII ^{es} Journées de Statistique*.

Rapports de recherche et publications internes

[45] C. Biernacki, G. Celeux, G. Govaert, « Strategies for Getting the Highest Likelihood in Mixture Models », *rapport de recherche n°4255*, Inria Rhône-Alpes, septembre 2001, `http://www.inria.fr/rrrt/rr-4255.html`.

[46] J. Diebolt, M. Garrido, S. Girard, « Le test ET : test d'adéquation d'un modèle central à une queue de distribution », *rapport de recherche*, Inria, 2001, RR-4170, `http://www.inria.fr/rrrt/rr-4170.html`.

[47] J. Diebolt, M. Garrido, C. Trottier, « A Bayesian Regularization Procedure for a Better Extremal Fit », *rapport de recherche*, Inria, 2001, RR-4211, `http://www.inria.fr/rrrt/rr-4211.html`.

[48] J. B. Durand, P. Gonçalvès, « Statistical Inference for Hidden Markov Tree Models and Application to Wavelet Trees », *rapport de recherche n°4248*, Inria Rhône-Alpes, 2001, `http://www.inria.fr/rrrt/rr-4248.html`.

[49] F. Forbes, C. Fraley, D. Georgian-Smith, D. Goldhaber, N. Peyrard, A. Raftery, « Region-Of-Interest Selection and Statistical Analysis of Dynamic Breast Magnetic Resonance Imaging Data », *rapport de recherche n°4249*, Inria Rhône-Alpes, 2001, `http://www.inria.fr/rrrt/rr-4249.html`.

[50] M. Garrido, « Régularisation bayésienne pour la loi de Weibull avec une loi a priori sur le paramètre de forme », *rapport de recherche*, mars 2001, EDF-Inria.

[51] A. Juditsky, S. Lambert-Lacroix, « On minimax density estimation on **R** », *rapport de recherche*, LMC, mai 2001, soumis à *J. of Bernoulli Soc.*

[52] N. Peyrard, « Convergence of MCEM and Related Algorithms for Hidden Markov Random Field », *rapport de recherche n°4146*, Inria Rhône-Alpes, 2001, `http://www.inria.fr/rrrt/rr-4146.html`.

Divers

[53] G. Celeux, F. Corset, « Modèles Graphiques appliqués à l'optimisation de maintenance pour le joint 1 d'une pompe primaire 900 MW », rapport de fin de contrat EDF-DER.

[54] F. Chouly, « Interest of the external field estimation of a Potts model for image segmentation », mémoire de 2^{eme} année - ENSIMAG, 2001.

[55] J. Diebolt, C. Bauby, M. Garrido, « Estimation bayésienne de la loi GPD, loi asymptotique des excès au-delà d'un seuil », rapport final de convention de recherche Inria-EDF, 2001.

[56] J. Gosme, *Représentations temps-fréquence diffusées : Propriétés, application et perspectives*, Mémoire, Univ. de Tech. de Troye, 2001.

[57] C. Goutte, C. Lavergne, Y. Vernaz, « Indices de sensibilité », 2001, rapport de contrat Inria Rhône-Alpes – CEA Cadarache.

[58] C. Goutte, C. Lavergne, « Analyse de sensibilité du code ACTIGE », 2001, rapport de contrat Inria Rhône-Alpes – CEA Cadarache.

[59] C. Goutte, C. Lavergne, « Décomposition de Sobol pour des distributions non uniformes des entrées », 2001, rapport de contrat Inria Rhône-Alpes – CEA Cadarache.

[60] C. Goutte, C. Lavergne, « Etudes du calcul des incertitudes par Circé », 2001, rapport de contrat Inria Rhône-Alpes – CEA Cadarache.

[61] A.-M. Koning, « Analyse discriminante sur tableaux de dissimilarités », Rapport de maîtrise, université de Caen, 2001.

[62] A. Regeasse, « Analyse des courbes intensité-temps issues d'examen par IRM en vue de l'aide au diagnostic du cancer du sein », rapport de DEA de Biostatistique, université de Montpellier 1, 2001.