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 [CR Inria, détachée de l'INPG]

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 depuis le 01/10/00]

Marcos Perreau-Guimares [boursier Inria]

Yann Vernaz [boursier Inria jusqu'au 31/10/00]

Chercheurs doctorants

Henri Bertholon [enseignant CNAM]

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

Franck Corset [boursier Inria depuis le 01/01/00]

Cécile Delhumeau [CHU de Grenoble]

Jean-Baptiste Durand [boursier MESR depuis le 1/10/99]

Myriam Garrido [boursière Inria depuis le 31/3/99]

Olivier Martin [boursier MESR, arrivé à IS2 le 01/07/00]

Nathalie Peyrard [boursière MESR depuis le 1/10/98]

Stagiaires longue durée

Véronique Equy [DEA MIMB, UJF]

Christelle Breuils [Université Paris VII, Denis Diderot]

Collaborateurs extérieurs

Christine Cans [médecin, association Rheops]

Jean Diebolt [DR CNRS au LMC-SMS jusqu'au 01/10/00, à l'Equipe d'Analyse et Mathématiques Appliquées de l'université de Marne-la-Vallée depuis le 01/10/00]

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

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, 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. Ainsi c'est notre culture sur les modèles à structure cachée qui nous a conduit 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 Diebolt, Jean-Baptiste Durand, Florence Forbes, 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 à la fois 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].

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

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

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.

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 soulève 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

[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.

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 *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

[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.

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 intérêt 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 [4] 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

Participants : Christian Lavergne, Yann Vernaz.

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 :

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

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

ii) d'un prédicteur linéaire η relié à l'espérance mathématique $E(Y) = \mu$ 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

Participants : Paulo Gonçalvès, Rudolf Riedi[1].

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.*

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 évolution est donnée par *la fonction de partition* :

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

1. Rice University, Houston (TX), USA.

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 l'égalité est stricte, 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, Christophe Biernacki, Christelle Breuils, Gilles Celeux, Franck Corset, Jean Diebolt, Cyril Goutte, Christian Lavergne, Myriam Garrido, Yann Vernaz.

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 on ne connaît pas sa valeur exacte mais 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 $x > 0$,

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

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 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, Véronique Equy, Jérome Fauconnier, Christian Lavergne, Claudine Robert, Paulo Gonçalvès, Jean-Baptiste Durand.

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). Un thème important de notre recherche concerne l'analyse des durées hospitalières de séjour.

Un autre sujet de recherche concerne l'étude du rythme cardiaque. Il s'agit de définir à partir d'enregistrements d'electrocardiogrammes (ECG), un ou plusieurs critères aidant à la classification des insuffisances cardiaques. Une piste envisagée repose sur l'identification de dépendances statistiques inter-échelle et à longue portée sur le rythme cardiaque (RR) et le temps de repolarisation ventriculaire (QT). Nous comptons utiliser des modèles de dépendance markovienne appliqués à la structure arborescente binaire issue de la décompositions en ondelettes de ces processus.

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.

Nous avons décidé de réaliser une toolbox Matlab prolongeant la SI-Toolbox de Lennart Ljung, conçu pour l'identification par des modèles linéaires paramétriques. L'interface de cette nouvelle boîte à outils sera très largement commune avec la SI-Toolbox.

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*, développés depuis quelques années dans le projet SIGMA2, utilisant 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[Jva95]. Outre les services d'identification proprement dite, on offre des moyens de valider une modélisation conduite avec une classe restreinte de modèles (par exemple, on peut tester si le modèle linéaire est ou non suffisant).

5.2 Le logiciel XEMGAUS

Participants : Christophe Biernacki, Gilles Celeux, Jean-Baptiste Durand, Gérard Govaert, Van Mô Dang.

Les mélanges multivariés gaussiens constituent un modèle de référence en analyse discriminante et en classification [MP00]. Ainsi, deux logiciels ont été récemment développés, MCLUST [9](`http://www.stat.washington.edu/fraley/mclust\_home.html`), écrit en Fortran et interfacé avec Splus, dédié à la classification hiérarchique et utilisant l'algorithme EM[DLR77], et le logiciel EMMIX (Peel et McLachlan, `http://www.maths.uq.edu.au/~gjm/emmix/emmix.html`), écrit en Fortran, qui utilise l'algorithme EM et sa version stochastique SEM[1] pour la classification et traite aussi le cas de l'analyse discriminante.

Nous avons développé, en Matlab, un logiciel concurrent à MCLUST et EMMIX : XEMGAUS. Il reprend nombre de leurs caractéristiques mais s'avère plus riche par l'intégration de nos recherches. Ainsi, il propose un grand nombre de modèles (28 au total) autorisant des variations sur la forme, l'orientation, le volume et la taille des classes, l'estimation peut se faire par différents algorithmes (EM, EM stochastique et EM classification) qui peuvent être enchaînés pour de meilleures performances, et le choix des meilleurs 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

[Jva95] A. JUDITSKY et al., «Nonlinear Black-Box Modelling in System Identification», *Automatica 31*, 12, 1995, p. 1725–1750.

[MP00] G. MCLACHLAN, K. PEEL, *Finite Mixture Models*, Wiley, New York, 2000.

[DLR77] A. DEMPSTER, N. LAIRD, D. RUBIN, «Maximum likelihood from incomplete data (with discussion)», *Journal of the Royal Statistical Society, Series B 39*, 1977, p. 1–38.

soi-même ses stratégies ou de s'en remettre à des choix par défaut. À terme, nous comptons le transcrire en C++ avec une interface Scilab.

Une première extension a été apportée au logiciel XEMGAUS. Elle a trait a l'estimation de paramètres pour les modèles de chaînes de Markov cachées. Les algorithmes suivants ont été développés pour l'identification : EM, EM à la Gibbs et algorithme de Viterbi. Le choix du nombre d'états cachés peut d'autre part être fait par une technique faisant appel à la validation croisée.

Une évolution essentielle est prévue : l'extension à d'autres types de mélange, comme les mélanges de Bernoulli pour les données binaires.

5.3 Le projet SEL

Participants : Marcos Perreau-Guimaraes, Claudine Robert, Bernard Ycart [2].

Marcos Perreau-Guimaraes a réalisé sous la direction de Bernard Ycart et en collaboration avec Claudine Robert un portail web (`http://www.inrialpes.fr/is2/`) de statistique en ligne à l'usage des enseignants en mathématiques du secondaire. 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.
  - *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 la EDF-DER, nous avons programmé un logiciel interactif en Matlab, interne à EDF, et intitulé EXTREMES.

2. Équipe Prisme, université René Descartes

Ce logiciel permet 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).

6 Résultats nouveaux

6.1 Modèles linéaires généralisés et hétéroscédasticité

6.1.1 Modèles additifs, autorégressifs et conditionnellement hétéroscédastiques

Participants : Christian Lavergne, Yann Vernaz.

Les modèles linéaires gaussiens ont dominé le développement de la modélisation des séries temporelles depuis plus de soixante ans. Cette phase a débuté avec les processus autorégressifs, pour se généraliser à la classe des modèles ARMA (*Auto-Regressive Moving Average*). Mais la classe des processus ARMA linéaires peut s'avérer inadaptée à certaines situations. On peut citer l'analyse des phénomènes monétaires et financiers dont les spécificités ne peuvent pas être prises en compte par une modélisation ARMA classique. Leur comportement est caractérisé par des dynamiques non linéaires et une volatilité (ou variabilité instantanée) marquée. Pour tenir compte de la volatilité, Engle[Eng82] propose une représentation autorégressive de la variance conditionnellement à son information passée. Cette classe de modèles est appelée arch (*Auto-Regressive Conditionnaly Heteroscedastic*). Le principe général proposé par Engle permet à la variance de dépendre de l'ensemble informationnel dont on dispose par une spécification où le carré des perturbations suit un processus autorégressif. L'idée générale pour obtenir de nouveaux modèles, dérivés du modèle ARCH, consiste à construire des modèles autorégressifs du type :

$$y_t = m(y_{t-1}, ...; \theta_0) + \sigma(y_{t-1}, ...; \theta_0) u_t \quad \text{pour } t = 1, 2, ... \tag{3}$$

où $m(.)$ et $\sigma^2(.) > 0$ sont les fonctions moyenne et variance conditionnelles et u_t est un bruit blanc indépendant des fonctions $m(.)$ et $\sigma(.)$. La stationnarité du processus y_t implique que le modèle n'est pas hétéroscédastique. L'équation (3) définit un modèle à temps discret avec des erreurs conditionnellement hétéroscédastiques (CH). Si la fonction moyenne est nulle, le processus est purement CH.

Nous proposons une méthode performante pour estimer les paramètres d'un modèle avec des erreurs CH lorsque la loi des erreurs est mal spécifiée. L'approche proposée s'inspire de

[Eng82] R. Engle, « Autoregresive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation », *Econometrika 54*, 1982, p. 987–1008.

la méthode de *quasi-vraisemblance*. Le concept de quasi-vraisemblance autorise une inférence statistique en ne retenant que les deux premiers moments et la fonction qui les relie. On exhibe alors des estimateurs construits à partir de la quasi-vraisemblance au sens de Hutton-Nelson ; la convergence et les propriétés asymptotiques de ces estimateurs ont été établies. Les avantages de l'algorithme sont sa simplicité de mise en œuvre, sa rapidité et sa stabilité numérique. La construction d'une fonction quasi-score permet également de disposer des tests classiques d'hypothèses : tests de quasi-Wald, quasi-score et rapport de quasi-vraisemblance.

Dans le contexte non paramétrique, les fonctions inconnues $m(.)$ et $\sigma(.)$ du modèle (3) sont estimées par la méthode les polynômes locaux. Nous proposons de plus une procédure adaptative du choix de fenêtre tirée de [LS97]. L'application de cet algorithme au modèle (3) permet d'obtenir une estimation optimale des fonctions inconnues en un point x. Les expérimentations numériques sur des données réelles et simulées confirment le bon comportement pratique des approches proposées[LV99], [7].

6.2 Problème de moindres carrés combinatoire

Participants : Anatoli Iouditski, Arkadi Nemirovski[3].

Des nombreux problèmes statistiques (tels que décodage de génome, détection multi-canaux, etc) nécessitent la résolution du problème d'optimisation suivant : étant donné une observation

$$y = Af_* + \sigma\xi$$

d'un vecteur f^* à n coordonnées, prenant les valeurs ± 1; la matrice A $m \times n$ donnée (ici ξ est un vecteur gaussian aléatoire), retrouver f_*.

Pour atteindre cet objectif, on peut utiliser la méthode du maximum de vraisemblance, laquelle dans notre cas devient la méthode des moindres carrés combinatoire, où l'estimation $\hat{f}$ du vrai « message » f_* est définie comme solution du problème d'optimisation :

$$\min_f \left\{ \|Af - y\|^2,\ f_i = \pm 1,\ \ i = 1..., n \right\}. \tag{4}$$

La difficulté cependant, est qu'aucune voie de calcul efficace pour résoudre le problème des moindres carrés combinatoire (4) n'est connue. Pour cette raison, nous avons étudié une méthode de résolution approximative de (4) « numériquement efficace », basée sur la procédure de relaxation semi-définie.

Nous avons conduit une série d'expériences numériques qui montrent que cette méthode donne des résultats comparables à ceux de la procédure des moindres carrés combinatoire.

3. Technion Haifa, Israel.

[LS97] O. Lepski, V. Spokoiny, « Optimal pointwise adaptative methos in nonparametric estimation », *Annals of Statistics 6*, 1997, p. 2512–2546.

[LV99] C. Lavergne, Y. Vernaz, « Estimation of Parametric Models with Conditional Heteroscedastic Errors », *rapport de recherche n° 3658*, Inria Rhône-Alpes, Grenoble, 1999.

6.3 Modèles à structure cachée

6.3.1 Accélération de l'algorithme EM pour les mélanges

Participants : Gilles Celeux, Stéphane Chrétien[4], Florence Forbes, Abdallah Mkhadri[5].

Souvent, l'algorithme EM converge lentement. Une des possibles raisons d'un tel comportement est le traitement simultané des paramètres à optimiser. Nous avons proposé [10] une version de l'algorithme EM pour l'estimation de mélanges de lois qui travaille composant par composant. Nous avons prouvé la convergence de cet algorithme en nous fondant sur son interprétation comme un algorithme proximal. Nous avons montré par des simulations que notre algorithme avait dans les situations de convergence lente un comportement meilleur que l'algorithme EM mais aussi que d'autres algorithmes tel que SAGE [FH94] par exemple, qui visent également à accélérer l'algorithme EM.

6.3.2 Stratégies d'obtention du maximim de vraisemblance pour les mélanges

Participants : Christophe Biernacki[6], Gilles Celeux, Gérard Govaert[7].

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 est souvent un problème difficile. Forts des nombreux algorithmes présents dans XEMGAUSS (EM, EM stochastique (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. Nous avons mené des expérimentations sur des données simulées et réelles en imposant un temps d'exécution fixé à l'avance. Bien qu'il soit difficile d'en tirer des conclusions définitives, il ressort de ces expériences, que l'utilisation d'un seul essai de l'algorithme EM est franchement mauvais. Il faut lui préférer l'emploi d'une combinaison répétée de CEM suivi de EM, ou mieux, une combinaison répétée de plusieurs exécutions courtes de l'algorithme EM suivi d'une exécution complète de ce même algorithme.

6.3.3 Approximation du champ moyen et segmentation d'images

Participants : Gilles Celeux, Florence Forbes, Nathalie Peyrard.

L'approximation du champ moyen est à l'origine une méthode d'approximation de la moyenne d'un champ de Markov. Elle est issue de la mécanique statistique où elle s'avère utile pour l'étude des phénomènes de transition de phases[Cha87]. Notre objectif est d'étudier

4. université de Besançon
5. université de Marrakech
6. université de Franche-Comté
7. université de technologie de Compiègne

[FH94] J. A. Fessler, A. Hero, « Space Alternating generalized expectation maximisation algorithm », *IEEE Trans. Signal Processing 42*, 1994, p. 2664–2677.

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

son utilisation, dans le cadre de la segmentation markovienne d'images, comme outil algorithmique pour éviter les calculs coûteux inhérents aux modèles de champs de Markov. L'idée est d'approximer les interactions entre pixels en négligeant les fluctuations : pour chacun des pixels, les pixels voisins sont supposés fixés à leur valeur moyenne. Cette méthode peut être vue comme une manière d'approximer un modèle markovien avec des interactions complexes par un système de variables indépendantes, beaucoup plus simple.

Dans le cadre de la segmentation d'images, nous nous sommes plus particulièrement intéressés à l'utilisation d'un tel outil pour l'algorithme EM. Pour des modèles markoviens, deux difficultés se présentent : le calcul de la fonction de partition a priori et celui des probabilités marginales a posteriori. Des approximations ont été proposées pour traiter ces étapes: pseudo-likelihood, MCMC. Zhang[Zha92], donne une solution heuristique utilisant des approximations de type champ moyen et obtient de bons résultats. Nous proposons une classe d'algorithmes fondés sur la généralisation du principe du champ moyen : dans le système indépendant approximant, les pixels voisins sont fixés à une constante, pas nécessairement égale à la valeur moyenne. En particulier, nous nous sommes intéressés à des méthodes de type champ modal ou champ simulé. Cette famille d'algorithmes contient la procédure proposée par Zhang, ainsi que la procédure (PPL)-EM de Qian et Titterington[QT91] et de nouveaux algorithmes dont celui du champ simulé qui apparaît très efficace sur les expérimentations que nous avons menées. De la sorte nous présentons sous une structure commune des algorithmes d'inspirations bien différentes et nous cherchons à les comparer en termes d'estimation des paramètres et de restauration de l'image. Un autre domaine d'application des techniques d'approximation de type champ moyen est la sélection de modèle (cf. 6.4.1).

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

Participants : Florence Forbes, Chris Fraley[8], Nathalie Peyrard, Adrian Raftery[7].

Dans le cadre d'une collaboration entre Toshiba MRI inc. à San Francisco, l'université de Washington à Seattle et IS2, le contexte est celui de l'Imagerie par Résonance Magnétique (IRM) pour la détection de cancers du sein. Cette technique permet potentiellement la différentiation des tumeurs malignes et des tissus sains mais a une spécificité réduite.

Dans une première approche, l'étude des caractéristiques dynamiques (courbes signal-temps) des tumeurs a été proposée pour améliorer la spécificité de l'IRM [Kuh00]. Une étape importante dans cette étude est la sélection d'une zone suspecte (ou ROI pour *Region-Of-Interest* en anglais). Il n'existe cependant pas encore de méthodes standardisées pour la sélection de telles zones et pour l'analyse des données d'IRM dynamique. Nous avons donc commencé une

8. Statistics Department, University of Washington, Seattle

[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.

[QT91] W. QIAN, D. TITTERINGTON, «Estimation of parameters in hidden Markov models», *Phil. Trans. R. Soc. Lond. 337*, 1991, p. 407–428.

[Kuh00] C. KUHL, «MRI of breast tumors», *European Radiology 10*, 2000, p. 46–58.

étude [40] qui propose une méthode de sélection fondée sur des techniques statistiques de classification multivariées, ainsi que des outils d'analyse des courbes pour les pixels sélectionnés.

Une seconde approche se base sur les travaux de R. Neugebauer, au cours de son DEA au printemps 2000 [50]. Il s'agit d'utiliser les principes de l'analyse discriminante pour permettre l'identification des classes de pixels, *i.e.* le diagnostic différentiel entre tissus cancéreux et tissus sains. Une telle analyse nécessite l'utilisation d'au moins une image d'apprentissage, où les pixels d'une tumeur maligne sont identifiés. Deux approches peuvent être distinguées selon la nature de l'image d'apprentissage. Dans le premier cas, l'image d'apprentissage et l'image étudiée ne sont pas issues de l'examen IRM du même patient. L'idéal serait en effet de pouvoir utiliser la coupe d'un ou plusieurs patients de référence, pour identifier les éventuelles zones cancéreuses des coupes d'autres patients. Cependant, ceci ne semble pas toujours facilement réalisable du fait du mode de recueil actuel des données IRM. Une autre approche s'avère néanmoins utile dans le suivi d'un patient lors d'un traitement médical par chimiothérapie. L'image d'apprentissage et l'image étudiée sont issues du même patient et du même examen IRM.

6.3.5 Extension des modèles de Markov cachés en reconnaissance de la parole

Participants : Florence Forbes, Alejandro Murua[9].

Nous avons commencé une étude sur un problème de reconnaissance de la la parole. Les questions soulevées sont du même ordre que celles abordées par J-B. Durand lors de son DEA[Dur99] et de sa thèse. Les chaînes de Markov cachées ont souvent été utilisées avec succès dans ce cadre mais elles présentent cependant un certain nombre de limitations. Dans [LM99], J. Li et A. Murua ont voulu tenter d'y remédier en proposant un modèle plus complexe, qui met notamment en jeu des dépendances entre variables non prises en compte habituellement. Cela donne lieu à des problèmes d'estimation des paramètres du modèle, résolus de manière essentiellement heuristique. L'objectif de notre collaboration est donc de voir si des techniques de type champ moyen, visant à se ramener à des cas de dépendances plus simples, pourraient s'appliquer et donner lieu à une procédure d'estimation mieux fondée statistiquement.

6.3.6 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 utilisée 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 vraisemblance du

9. Statistics Department, University of Washington, Seattle

[Dur99] J. DURAND, *Reconnaissance statistique de trajectoires par modèles de Markov cachés*, Mémoire, Université Joseph Fourier, Institut National Polytechnique de Grenoble, 1999.

[LM99] J. LI, A. MURUA, « A 2D extended HMM for speech recognition », *in : Proceedings of the International Conference on Acoustics, Speech and Signal Processing ICASSP*, 1999.

modèle par des techniques d'*half sampling* qui sont un cas particulier de la validation croisée. Dans le principe, cela consiste à diviser l'échantillon en deux, puis à estimer les paramètres sur une partie et à calculer la vraisemblance du modèle sur l'autre. En inversant le rôle des deux parties, 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 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. Les premières expérimentations menées sont encourageantes et semblent indiquer une certaine supériorité de la procédure se fondant sur un partionnement alternatif de l'échantillon.

6.3.7 Indexation d'images

Participants : Christophe Biernacki[10], Anne Guérin-Dugué, Jeanny Hérault[11].

Cette étude a été initiée par Christophe Biernacki lors de son stage post-doctoral en 1999, en collaboration entre Movi et IS2. La modélisation des distributions de caractéristiques extraites dans les images est ici appliquée au contexte de l'indexation des images par le contenu. L'objectif est de retrouver dans une base d'images, les images les plus « similaires » à une image requête suivant la similarité entre les distributions d'information de chrominance seule, ou combinée avec l'information de luminance, par le biais de l'estimation d'une caractéristique locale d'orientation. Le principe retenu est d'indexer l'image par les paramètres optimaux (au sens du critère ICL [8]) du mélange de gaussiennes modélisant la distribution globale des caractéristiques. L'appariement se réalise en maximisant la vraisemblance du jeu de paramètres d'une image de la base sachant la distribution empirique de l'image requête. Plusieurs espaces chromatiques ont été testés. Les meilleurs résultats ont été obtenus avec l'espace conduisant en moyenne à une modélisation la plus parcimonieuse de la distribution chromatique. Cet espace de représentation chromatique est issu des travaux de recherche effectués au Laboratoire des Images et des Signaux (LIS-INPG-UJF) sur la vision humaine des couleurs.

Cette approche a été étendue avec succès dans deux directions. La première extension concerne la modélisation des distributions chromatiques spatialement localisées dans l'image. En effet, l'organisation spatiale des modes détectés est perdue avec la seule modélisation de la distribution globale. Cette organisation spatiale peut être retrouvée par la suite en effectuant une étape de classification spatiale des modes détectés. L'étape de modélisation concerne alors la distribution conjointe spatio-chromatique [41]. L'autre extension est relative à la fusion des informations de chrominance et de luminance (orientation). Cette fusion est réalisée sur la vraisemblance en considérant les deux informations indépendantes [49].

10. Université de Franche-Comté
11. Université Joseph Fourier

6.4 Choix de modèles en discrimination et classification automatique

6.4.1 Sélection de modèle pour les champs de Markov caché

Participants : Gilles Celeux, Florence Forbes, Nathalie Peyrard, Adrian Raftery[4].

Pour des modèles de mélanges indépendants, il existe plusieurs critères pour sélectionner le nombre de classes présentes. Le critère BIC, ainsi que le critère ICL [8] nécessitent le calcul de l'estimateur du maximum de vraisemblance des paramètres et du mode de la distribution a posteriori des données cachées. Lorsque le modèle de mélange est défini par un champ de Markov caché, ces deux quantités ne peuvent plus être calculées exactement. Nous étudions des approximations basées sur des méthodes de type champ moyen ou pseudo-vraisemblance. Celles-ci permettent d'éviter la lourdeur des méthodes de type chaînes de Markov de Monte-Carlo (MCMC) [27].

6.4.2 Combinaison de modèles en analyse discriminante

Participants : Isabel Brito, Gilles Celeux, Ana Maria Sousa Ferreira[12].

Ce thème constitue le sujet des thèses que préparent Isabel Brito pour les méthodes de discrimination sur variables quantitatives et Ana Maria Sousa Ferreira qui considère des modèles de discrimination sur variables qualitatives. 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 exploré les possibilités du couplage hiérarchique pour des problèmes à plus de deux classes. Cela consiste à voir le problème comme une suite de problèmes à deux classes qui ne seront pas nécessairement séparées par un même modèle. À chaque étape de la construction hiérarchique, les deux classes à séparer et le modèle sont choisis par validation croisée. Des expérimentations montrent l'intérêt de la méthode dans les cas où certaines classes sont faciles à séparer et d'autres non. Ce type de technique de couplage hiérarchique donne de bonnes performances avec une lisibilité accrue des règles de décision construites [24]. Dans le cadre qualitatif où travaille Ana Maria Sousa Ferreira, le couplage hiérarchique a aussi été employé. Mais ici la construction de l'arbre se fait en utilisant la distance de Matusita entre les distributions de probabilité de chaque classe. De plus, à chaque nœud, le modèle multinomial complet et le modèle d'indépendance conditionnelle sont combinés de manière optimale à partir du facteur de Bayes dont on calcule la valeur exacte dans un cadre non informatif [29].

6.4.3 Analyse discriminante sur tableaux de dissimilarités

Participants : Gilles Celeux, Anne Guérin-Dugué.

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

12. université de Lisbonne

L'objectif de ces travaux est de proposer une nouvelle approche aux deux stratégies usuellement suivies. La première voie est d'utiliser un outils de discrimination basé sur l'algorithme des « K plus proches voisins » (méthode coûteuse, mal adaptée aux classes non sphérique, mais bien adaptée à des classes non connexes). La seconde voie est de transformer ce problème non métrique en un problème métrique par le biais d'un algorithme de prolongement euclidien (*MultiDimensionnal Scaling*) avec un risque de distorsions importantes.

Notre approche s'inspire des modèles gaussiens. Une famille de trois règles de complexité croissante a été proposée, pour s'adapter aux diverses structures de données. Le principe est d'estimer à partir de statistiques simples (moyenne, variance) sur les tableaux de dissimilarités, des quantités sensibles à la position et à la forme des classes si le prolongement euclidien de celles-ci était connu. L'adaptation aux données s'effectue à l'aide de paramètres d'ajustement (un par classe) choisis de façon à minimiser le taux d'erreur estimé par validation croisée [34]. Une extension aux tableaux creux, où toutes les dissimilarités entre objets ne sont pas connues, a été développée. C'est un cas pratique important vu la tendance à traiter d'énormes bases de données susceptibles de présenter de nombreuses données manquantes. La méthode a été validée sur des données artificielles et réelles (classification de protéines) avec des résultats comparables à une classification par « K plus proches voisins », un temps de calcul beaucoup plus court, et une grande robustesse à la structure creuse de la matrice de dissimilarités jusqu'à 40-50% de valeurs manquantes.

6.5 Modèles de fiabilité industrielle

6.5.1 Un modèle de vieillissement

Participants : Henri Bertholon, Gilles Celeux.

Nous avons poursuivi notre analyse du modèle de viellissement apparenté à la loi de Weibull mais qui introduit un instant de début de vieillissement t_0 strictement positif [22]. L'intérêt d'un tel modèle est de pouvoir anticiper le début de vieillissement afin d'optimiser une politique de maintenance préventive. Cette année, nous avons étendu notre test optimal de l'existence d'un vieillissement obéissant à notre modèle au cas où le paramètre d'échelle de la loi exponentielle qui décrit l'absence de vieillissement était inconnu. Par ailleurs, nous avons conçu un algorithme EM conditionnellement à t_0 pour estimer les paramètres du modèle. Il est possible que l'emploi de cet algorithme nous permette une estimation raisonnable du paramètre de forme gouvernant le vieillissement que nous supposions fixé dans la version antérieure, mais des experimentations doivent être faites.

6.5.2 Étude de problèmes de fiabilité dans un contexte de données doublement censurées

Participants : Gilles Celeux, Christian Lavergne, Yann Vernaz.

Cette recherche traite de problèmes issus d'un partenariat avec le département « Sûreté de fonctionnement, Diagnostic, Maintenance » de EDF-DER.

On s'intéresse à la modélisation de la durée de vie d'un matériel réparable à partir de données doublement censurées. Aucune date de défaillance n'est observée. Ce cas de figure est

caractéristique lorsqu'on effectue des contrôles réguliers d'un matériel. En effet, les données disponibles sont alors les dates de contrôles pour lesquelles on a le nombre de censures à droite (pas de défaillance) et le nombre de censures à gauche (on relève des défaillances mais on ne connaît pas la date exacte des défaillances).

La loi de Weibull fournit une modélisation souvent appropriée pour tenir compte du vieillissement d'un matériel. Cependant, eu égard à la nature des données, l'estimation des paramètres de la distribution de Weibull s'avère difficile voire impraticable. Nous étudions la performance des estimations des paramètres de la loi de Weibull dans le contexte d'une inférence directe par la méthode du maximum de vraisemblance puis par une approche bayésienne non informative. Les résultats numériques font apparaître une supériorité nette de l'approche bayésienne, mais ils ne sont pas toujours satisfaisants et nous amènent à proposer une alternative à la modélisation de Weibull. La méthode consiste à considérer les censures à gauche (peu nombreuses) comme la réalisation d'un processus de Poisson puis de plonger le processus de Poisson dans un modèle GLM. On traite alors une défaillance comme un événement rare et on modélise le vieillissement par palier. Cette méthodologie fournit des réponses simples et fiables aux questions que l'on se pose (par exemple l'existence d'un vieillissement) à l'aide des tests d'hypothèse classiques issues de la théorie des modèles GLM [38].

6.5.3 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.

Une étude bibliographique préliminaire nous a permis de définir des principes pour le choix des variables notamment à travers un article[Hø96]) et d'un projet EDF[CL]. Actuellement, Nous en sommes à la constitution du graphe, c'est-à-dire à la définition des influences directes entre les variables en jeu et à l'évalutation des probabilités attachées aux arêtes ainsi constituées. Ce travail se fait avec un groupe de cinq experts EDF.

Par ailleurs, en collaboration avec Stéphane Chrétien de l'université de Franche-Comté, Franck Corset a étudié le comportement asymptotique d'un minimiseur et de l'optimum pour un problème de plus court chemin dans un graphe où les arêtes sont pondérées par des poids stochastiques. La convergence du minimiseur et un théorème central limite pour l'optimum sont prouvés en analysant le problème du plus court chemin stochastique comme une solution d'un problème linéaire [PS98] où les contraintes sont déterministes. Ce type de résultat pourra

[Hø96] S. Højsgaard, « Learning structures from data and experts », *Mathematics and Computers in Simulation 42*, 1996, p. 143–152.

[CL] C. Chatelain, A. Lannoy, « Une application de la technique du réseau bayésien à l'évolution des couts de maintenance », projet DER-EDF 2000.

[PS98] C. Papadimitriou, K. Steiglitz, *Combinatorial Optimization : algorithms and complexity*, Mi-

amener à déterminer des intervalles de confiance pour la longueur du pire chemin (celui qui conduira le plus rapidement à la ruine du système) dans les modèles graphiques qui nous occupent.

6.5.4 Modélisation d'un changement de comportement de maintenance

Participants : Christelle Breuils, 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 sommes intéressés au problème suivant : 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. Les résultats expérimentaux sont encourageants, mais une validation par simulation de Monte-Carlo pour des situations réalistes reste à faire.

6.5.5 Modélisation et estimation de queues de distributions

Participants : Jean Diebolt, Myriam Garrido.

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 au-delà de la dernière observation.

Lors d'un premier contrat[DDEA97], il est ressorti que la méthode ET pouvait être un moyen simple de réaliser l'estimation de ces quantiles. Un deuxième contrat[GDD98] 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[DDT99] 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.

---

neola, NY : Dover publications, 1998, ch. 3.

[DDEA97] J. Diebolt, V. Durbec, M. A. El Aroui, «Modélisation de queues de distributions et estimation de quantiles extrêmes», Rapport final de convention de recherche Inria-EDF, 1997.

[GDD98] S. Girard, J. Diebolt, V. Durbec, «Modélisation de queues de distributions et estimation de quantiles extrêmes(2) », Rapport final de convention de recherche Inria-EDF, 1998.

[DDT99] J. Diebolt, V. Durbec, C. Trottier, « Régularisation de distributions pour une meilleure adéquation extrême», 1999, rapport final de convention de recherche Inria-EDF.

Le présent contrat a permis de continuer l'étude du test d'adéquation à la queue de distribution précédemment proposé. Nous étudions 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 obtiendra 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.

6.6 Statistique biomédicale

Participants : Christine Cans, Gilles Celeux, Cécile Delhumeau, Jérôme Fauconnier, Christian Lavergne, Claudine Robert, Veronique Équy.

6.6.1 Surveillance de l'infirmité motrice d'origine cérébrale en Europe

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 créer un réseau de registres de morbidité basé sur l'étude des enfants atteints de CP. Une base de données exploitable commune à 14 centres européens regroupant des caractéristiques médico-sociales comporte les cas d'enfants atteints de CP nés entre 1975 et 1990, et des informations sur quelques indicateurs périnatals relatifs à la population de référence pour chacun des centres. Elle devrait permettre 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 les centres selon des indicateurs périnatals observés dans la population générale où est étudiée la CP : 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 bébés pesant moins de 1 500g. Nous cherchons à savoir s'il faut distinguer les centres entre eux et s'il faut également différencier les années et le sexe. Nous attaquons ce problème par la régression logistique. Les résultats font apparaître des différences notables entre centres, reflet de politiques de soins et de pratiques médicales différentes.

6.6.2 Analyse des durées de séjour du CHU de Grenoble

Participants : Gilles Celeux, Cécile Delhumeau, Jérôme Fauconnier.

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.

Pour ce faire, nous avons mené une comparaison de la répartition des quartiles (quantiles à 25%) des DS des GHM grenoblois par rapport à ceux de la base nationale afin de repérer les GHM grenoblois atypiques en termes de DS et d'essayer de les classer selon leurs distributions. Nous avons modélisé ces distributions de quantiles de DS par une analyse en composante principale (ACP) et par un modèle de mélange de lois normales qui utilise le logiciel XEMGAUS. Tenant compte des interprétations médicales, une structuration en trois groupes dont un groupe de 114 GHM grenoblois «mauvais». Vingt-deux GHM symptomatiques de ce groupe ont été sélectionnés pour comparer leurs distributions à Grenoble et sur la base nationale afin de caractériser leurs différences et notamment de voir si l'allongement des DS des GHM grenoblois est lié à une prise en charge différente des patients.

6.6.3 Mesure de la clarté nucale fœtale : harmonisation des pratiques

Participants : Véronique Équy, Christian Lavergne.

La mesure de la clarté nucale fœtale lors de l'échographie de la fin du premier trimestre de la grossesse est un outil majeur de dépistage en diagnostic ante-natal, en particulier de la trisomie 21. Dans le cadre d'un DEA de modélisation, nous avons réalisé une étude rétrospective cherchant à modéliser la clarté nucale en fonction d'autres variables enregistrées lors de l'échographie. Nous avons pu mettre en évidence un effet praticien très significatif sur cette mesure. Nous présentons un protocole d'étude pour tenter de remédier à ce problème majeur. Nous envisageons d'une part une approche méthodologique et d'autre part, d'utiliser un outil de contrôle et de mesure automatique de la clarté nucale, qui s'il permettait de s'affranchir de ce biais, pourrait permettre d'instaurer un contrôle de qualité des mesures échographiques. Enfin, nous suggérons une nouvelle méthode de dépistage de la trisomie 21 basée sur l'étude des résidus élevés du modèle. Ce travail est l'objet de la thèse de sciences ingénierie du vivant que vient d'entamer Véronique Équy.

6.7 Ondelettes et lois d'échelle

6.7.1 Statistiques au second ordre de l'estimateur

Participants : Paulo Gonçalvès, Rudolf Riedi[13], Richard Baraniuk[14].

On sait que les coefficients d'ondelette de la décomposition d'un mouvement brownien fractionnaire (mbf) de paramètre H:
(i) sont stationnaires (au second ordre) à chaque échelle,
(ii) sont faiblement corrélés $EC_{n,k}\,C_{n',k'} \sim \sigma_n \delta_{n,n'}\,\delta_{k,k'}$,
(iii) suivent la loi de scaling suivante $d_{n,k} \stackrel{\mathrm{d}}{=} 2^{nH} d_{0,k}$.

13. Rice university, Houston (TX), USA.
14. université de Lisbonne

Il est dans ces conditions possible d'obtenir une forme explicite du spectre de singularité $f(\alpha)$ de ce processus

$$f(\alpha) \begin{cases} = & -\infty & \alpha < H \\ = & 0 & \alpha = H \\ = & H - \alpha & \alpha > H. \end{cases}$$

Ce spectre est dit *dégénéré* ou *trivial*. Dans [GRB98], nous avions établi des résultats asymptotiques sur les statistiques au premier et second ordre de l'estimateur empirique du spectre de Legendre.

Nous reprenons dans [32] une étude analogue mais portant cette fois-ci sur un processus dans lequel cohabitent plusieurs lois d'échelle de natures différentes (i.e. un processus offrant un spectre non dégénéré). Le processus que nous étudions combine mouvement brownien fractionnaire $B_H(t)$ et mesure multifractale $\mathcal{M}(t)$ à travers la loi de composition suivante

$$\mathcal{B}_H(t) : = B_H(\mathcal{M}(t)).$$

Le spectre de singularité de ce processus composé vaut alors

$$f_{\mathcal{B}}(\alpha) = f_{\mathcal{M}}(\alpha/H),$$

et sa richesse provient exclusivement de celle du spectre $f_{\mathcal{M}}(\alpha)$ de la mesure. Le mbf n'est dans ce schéma qu'un vecteur qui sert de support au développement d'une infinité de lois d'échelle.

Comme pour le mbf seul, nous avons donc montré que les coefficients en ondelettes résultant de la décomposition de $\mathcal{B}_H(t)$ sont identiquement distribués selon la loi

$$C_{n,k} \stackrel{d}{=} (W_n)^H \, C_{0,0},$$

où W_n est une variable aléatoire ne dépendant que de l'indice d'échelle n; stationnaires au second ordre, la fonction de covariance ne dépendant des positions k et k' que par leur différence et à décorrélation rapide. Nous travaillons actuellement à la détermination des statistiques au premier et second ordre de l'estimations de $f_{\mathcal{B}}(\alpha)$.

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

Participants : Paulo Gonçalvès, Rudolf Riedi[15].

S'agissant d'un mbf (processus gaussien), on sait que les coefficients d'ondelette (eux même gaussiens) élevés à la puissance q, avec $q < -1$, suivent une loi $\gamma-$stable (d'espérance infinie), et l'estimateur empirique $S^n(q)$ 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

15. Rice university, Houston (TX), USA.

[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.

quels ordres de q les moments de la v.a $|C_{n,k}|$ existent. Nous avons donc proposé dans [33], un test simple permettant de déterminer le domaine d'existence (q_{min}, q_{max}) de la fonction $S^n(q)$. Pour une variable aléatoire continue $\mathbf{x}$, ce test repose sur la régularité locale de sa fonction caractéristique $\chi(y)$ au voisinage de zéro. Nous procédons suivant deux étapes:
Étape 1. À partir de N réalisations i.i.d. de la variable aléatoire $\mathbf{x}$, nous construisons l'estimateur empirique de la fonction caractéristique

$$\chi(y_j) \;=\; \frac{1}{N}\sum_{i=1}^{N}\exp(ix_iy_j).$$

N'étant intéressés que par le comportement de cette fonction au voisinage de l'origine, nous choisissons une grille d'échantillonnage $y_j = j.\delta y$, localisée autour de zéro avec un pas δy conditionné par les données initiales x_i.
Étape 2. La régularité ponctuelle (hölderienne) $\alpha(t)$ d'une fonction peut se définir à partir de sa décomposition en ondelettes, par le taux de décroissance des supremum des coefficients dans un voisinage de t. Réciproquement, on peut estimer cette régularité $\alpha(t)$ en mesurant le taux de décroissance (log-linéaire) des maxima locaux des coefficients d'ondelette autour du point t. Nous procédons donc à la décomposition en ondelettes de la fonction estimée $\chi(y)$, et mesurons la régularité de celle-ci à l'origine $y = 0$. La partie entière de $\alpha(0)$, ainsi obtenue, nous indique l'ordre maximal q_{max} au delà duquel les moments de la variable aléatoire $\mathbf{x}$ n'existent pas.

Pour obtenir la borne inférieure q_{min} de $S^n(q)$, nous procédons de la même façon avec la nouvelle variable aléatoire $\mathbf{z} = \mathbf{x}^{-1}$. Un sous-produit de cette procédure permet aussi d'estimer les paramètres de certaines classes de distributions (e.g. γ-stable, β-distributions,...).

6.7.3 Synthèse d'ondelettes

Participants : Paulo Gonçalvès, Claude Lemaréchal[16], François Oustry[17], Coralie Triadou[18].

Nous étudions en collaboration avec le projet Numopt une procédure d'optimisation pour la synthèse d'ondelettes (orthogonales en un certain sens défini dans [GA97]) destinées à l'estimation de la régularité ponctuelle de processus (cf. rapport d'activité 2000 Numopt, *Synthèse d'ondelettes*).

7 Contrats industriels (nationaux, européens et internationaux)

7.1 Retour d'expérience de constituants de pompes

Participants : Christelle Breuils, Gilles Celeux, Franck Corset.

Ce contrat de type CERD avec le département «Surveillance, Diagnostic, Maintenance» de

16. Permanent Numopt
17. Permanent Numopt
18. Stagiaire Numopt de deuxième année Ensimag.

[GA97] P. Gonçalvès, P. Abry, «Multiple-Window Wavelet Transform and Local Scaling Exponent Estimation», *in: IEEE, Int. Conf. on Acoust. Speech and Signal Proc.*, Munich (Germany), 1997.

la DER-EDF concernait l'analyse statistique du retour d'expérience des constituants de pompes primaires afin, en particulier, de juger de leur vieillissement. Cette étude faisait suite à une étude de même type effectuée l'an dernier. Cette année nous nous sommes concentrés sur des problèmes d'analyse discriminante pour déceler les facteurs suceptibles de provoquer un rebut des matériels en jeu et nous avons proposé un modèle permettant de caractériser la loi de duré de vie des matériels en se débarassant de l'effet néfaste des rebuts ordonnés à tort par précaution (cf. 6.5.4 et [43]).

7.2 Contrat EDF sur les queues de distribution de probabilité

Participants : Jean Diebolt, Myriam Garrido.

Ce contrat de type CERD 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_{α_n} défini par :

$$P(X > q_{\alpha_n}) = \alpha_n, \quad q_{\alpha_n} < 1/n.$$

Cette étude prolonge le travail des deux 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. Notons que suite à ces contrats, Jean Diebolt travaille avec Philippe Barbe (CNRS, Évry et université de Yale) sur l'analyse du processus empirique des excès.

7.3 Contrat DEA (Cadarache): Étude d'incertitudes et de sensibilité

Participants : Christian Lavergne, Yann Vernaz.

Nous avons démarré une collaboration sur 3 ans 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, les scénarios d'accidents graves.

En 2000, nous avons développé pour le CEA des outils permettant d'évaluer la sensibilité des paramètres d'entrée sur un système non linèaire et non monotone. Les travaux effectués à ce jour sont très encourageants et nous mettons au point une nouvelle méthode de sensibilité basée sur une décomposition en ondelettes. De plus, un travail important a été fait sur l'analyse discriminante des données issues des codes de calcul du DEA.

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, M. Garrido, P. Gonçalvès, C. Goutte 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).

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.

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. Avec I. Brito, ils ont participé à l'atelier sur les problèmes de validation en classification automatique qui s'est tenu à la maison du Portugal à la cité universitaire de Paris.

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) sont nos collaborateurs au Portugal, et avec eux nous mènerons une étude sur *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).

Amérique du Nord

Le projet IS2 poursuit sa collaboration avec le département de statistique de l'université de Washington à Seattle. N. Peyrard et F. Forbes ont effectué un séjour de trois mois dans ce département et G. Celeux un séjour d'un mois. À cette occasion, ils ont participé au groupe de travail « Model-Based Clustering and Applications » organisé par A. Raftery.

9 Diffusion de résultats

9.1 Animation de la communauté scientifique

G. Celeux et J. Diebolt ont organisé une Journée Inria Rhône-Alpes sur les quantiles extrêmes en mars 2000 à laquelle M. Garrido a contribué.

J. Diebolt et G. Celeux ont participé au groupe de travail MC.Cube du Crest, Ensae. Ce groupe de travail réunit, dans la mesure du possible, tous les chercheurs intéressés par le thème des méthodes de contrôle de la convergence vers la stationnarité des chaînes de Markov engendrées par les algorithmes MCMC. Il se réunit approximativement une fois par mois, à l'initiative de Christian Robert (Crest).

G. Celeux, J. Diebolt, F. Forbes, C. Lavergne et N. Peyrard participent à un groupe de travail sur le thème de la modélisation spatiale réunissant des chercheurs des laboratoires grenoblois LMC-SMS et Labsad.

P. Gonçalvès est responsable avec C. Doncarli (IrCyn, ECN) du groupe de travail *Analyse et Décision en Signal* du GDR-PRC ISIS (CNRS). C. Lavergne et F. Forbes ont fait un exposé à la journée plénière de ce groupe de travail. 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 cordonnateur avec P. Abry (ENS-Lyon) et J. Lévy-Véhel (projet Fractales de l'UR de Rocquencourt) d'un volume *Fractals et lois d'échelle* 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 Joseph-Fourier / Ensimag, Grenoble (12 heures par an) sur les tests d'adéquation non paramétriques, et sur la théorie asymptotique de l'estimation pour les processus autorégressifs.

P. Gonçalvès a assuré en janvier 2000, un cours de 8 heures sur *Ondelettes et Fractales* au DEA de « Traitement des Images et du Signal » de l'ENSEA (Université de Cergy-Pontoise).

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 participé à la session invitée sur « model-based clustering » organisée par A. Raftery lors du *Joint Statitistical Meeting* à Indianapolis en août.

N. Peyrard a participé à la *First European Conference on Spatial and Computational Statistics*, à Ambleside, Grande-Bretagne en septembre.

G. Celeux, C. Lavergne, F. Forbes, N. Peyrard et Y. Vernaz ont participé aux XXXIIèmes journées de statistique de la SfdS, à Fès (Maroc) en mai 2000.

I. Brito a participé à IFCS 2000, rencontre internationale des sociétés de classification, à Namur en juillet.

P. Gonçalvès a été conférencier invité à *International Conference on Telecommunications*, Acapulco (Mexique), Mai 2000.

H. Bertholon et M. Garrido ont participé à MMR'2000 *Second Internationnal Conference on Mathematical Methods in Reliability* à Bordeaux en juillet.

10 Bibliographie

Ouvrages et articles de référence de l'équipe

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

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

[3] M. El-Aroui, C. Lavergne, «Generalized linear models in software reliability, parametric and semi-parametric approaches», *IEEE Trans. on Reliability 43*, 1996, p. 463–471.

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

[5] 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.

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

Thèses et habilitations à diriger des recherches

[7] Y. Vernaz, *Contributions à l'estimation de modèles conditionnellement hétéroscédastiques et à l'étude de problèmes de fiabilité dans un contexte de données doublement censurées*, thèse de doctorat, Université Joseph Fourier, 2000.

Articles et chapitres de livre

[8] C. Biernacki, G. Celeux, G. Govaert, «Assessing a mixture modelfor clustering with the integrated completed likelihood», *IEEE Trans. on PAMI 20*, 2000, p. 267–272.

[9] C. Cans, P. Guillem, C. Lavergne, «Comment calculer l'exhautivité d'un registre de morbidité? L'exemple du registre des handicaps de l'enfant et observatoire périnatal de l'Isère», *Revue d'Épidémiologie et de Santé Publique 48*, 2000, p. 41–51.

[10] G. Celeux, F. Forbes, A. Mkhadri, S. Chrétien, «A Component-Wise EM algorithm for Mixtures», *Journal of Computational and Graphical Statistics*, 2000, à paraître.

[11] G. Celeux, M. Hurn, C. Robert, «Computational and inferential difficulties with mixture posterior distributions», *Journal of the American Statistical Association 95*, 2000, p. 957–970.

[12] J. Diebolt, V. Durbec, M. El Aroui, B. Villain, « Estimation of extreme quantiles: empirical tools for methods assessment and comparison», *International Journal of Reliability, Quality and Safety Engineering 7(1)*, 2000, p. 75–94.

[13] J. DIEBOLT, J. ZUBER, «On testing the goodness-of-fit of a nonlinear heteroscedastic regression model», *Communications in Statistics*, 2000, à paraître.

[14] P. FLANDRIN, P. GONÇALVÈS, P. ABRY, «Analyses en ondelettes et lois d'échelle», *in : Fractals et lois d'échelle, Information-Commande-Communication*, Hermes Science, 2001, à paraître.

[15] O. FRANÇOIS, C. LAVERGNE, «"Experimental Design for Evolutionary Algorithms: a Statistical Perspective», *IEEE Transactions on Evolutionary Computation*, 2000, à paraître.

[16] S. C. GROUP, «Surveillance of Cerebral Palsy in Europe : A European collaboration of cerebral palsy surveys and registers», *Developmental Medicine and Child Neurology*, 2000, à paraître.

[17] M. HRISTACHE, A. JUDITSKY, V. SPOKOINY, «Direct Estimation of the Index Coefficients in a Single-index Model», *Ann. of Stats*, 2000, à paraître.

[18] C. LAVERGNE, C. TROTTIER, «Sur l'estimation dans les modèles linéaires généralisés à effets aléatoires», *Revue de Statistique Appliquée XLVIII (1)*, 2000, p. 45–63.

[19] A. NAZIN, A. JUDITSKY, «On minimax approach to nonparametric adaptive control», *Int. J. of Adaptive Contr. and Signal Proc.*, 2000, à paraître.

[20] 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.

[21] C. TROTTIER, «A quasi-marginal approach in generalized linear mixed models», *Statistics 33*, 2000, p. 291–308.

Communications à des congrès, colloques, etc.

[22] H. BERTHOLON, «A change point Ageing model», *in : MMR'2000 abstracts' book*, p. 199–201, Bordeaux, France, juillet 2000.

[23] C. BIERNACKI, G. CELEUX, G. GOVAERT, «Stratégies algorithmiques pour l'estimation des mélanges», *in : XXXIIèmes journées de statistique*, Fès - Maroc, 15 -19 mai 2000.

[24] I. BRITO, G. CELEUX, «Discriminant analysis by Hierarchical coupling in EDDA context», *in : Data Analysis, Classification and related Methods*, Springer, p. 175–180, Namur, 11-14 juillet 2000.

[25] G. CELEUX, F. FORBES, N. PEYRARD, «Critères pour la sélection d'un modèle de champ de Markov caché», *in : XXXIIèmes journées de statistique*, Fès - Maroc, 15 -19 mai 2000.

[26] G. CELEUX, F. FORBES, N. PEYRARD, «Mean field approximation methods for parameter estimation and model selection in hidden Markov models for images», *in : Model-Based Clustering and Applications Workshop*, p. 243–280, Seattle - USA, 24 -28 juillet, 2000.

[27] 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.

[28] J. DIEBOLT, «Un test d'adéquation pour modèles de régression non linéaire», *in : Goodness of Fit 2000*, Université Paris-5, France, mai 2000.

[29] A. Ferreira, G. Celeux, H. Bacelar, «Discrete discriminant analysis: the performance of combining models by a hierarchical coupling approach», *in : Data Analysis, Classification and related Methods*, p. 181–186, Namur, 11-14 juillet 2000.

[30] M. Garrido, J. Diebolt, «The ET test, a goodness-of-fit test for the distribution tail», *in : MMR'2000 abstracts' book*, p. 427–430, Bordeaux, France, juillet 2000.

[31] S. Girard, J. Diebolt, «Consequence of the Picklands approximation on the extreme quantile approximation», *in : MMR'2000 abstracts' book*, p. 459–462, Bordeaux, France, juillet 2000.

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

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

[34] A. Guérin-Dugué, G. Celeux, «Discriminant Analysis on Dissimilarity Data : A New Fast Gaussian-like Algorithm», *in : AISTATS*, 2001.

[35] C. Lavergne, Y. Vernaz, «Application d'une procédure de choix de fenêtre adaptative pour un modèle CH», *in : XXXIIèmes journées de la SFdS*, Fès, Maroc, 15-19 mai 2000.

[36] C. Trottier, J. Diebolt, « Adjusting density functions for a better extremal fit », *in : MMR'2000 abstracts' book*, p. 995–999, Bordeaux, France, juillet 2000.

Rapports de recherche et publications internes

[37] H. Berthelon, G. Celeux, A. Lannoy, «Évaluation de la constance du taux de défaillance», *rapport de recherche*, janvier 2000, EDF-DER.

[38] G. Celeux, C. Lavergne, Y. Vernaz, « Assessing Material Aging from Doubly Censored Data: Weibull Distribution vs. Poisson Process », *rapport de recherche n° 3857*, Inria Rhône-Alpes, 2000.

[39] G. Celeux, M. Persoz, V. Venturini, «Bayesian Modeling of PWR Vessel Flow Distributions taking into account indications», *rapport de recherche*, avril 2000, EDF-DER.

[40] F. Forbes, N. Peyrard, C. Fraley, A. Raftery, «Region-Of-Interest selection and dynamic breast magnetic Resonance data analysis via multivariate and spatial statistical segmentation methods», *rapport de recherche*, octobre 2000, draft.

Divers

[41] S. Blanck, *Modélisation statistique de la chrominance dans les images en couleur : Application à l'indexation de base d'images*, Mémoire, université de Franche Comté, 2000.

[42] C. Breuils, G. Celeux, F. Corset, «Analyse statistique des durées de vie pour les pompes primaires 900 et 1300 MW», 2000, rapport final de contrat Inria Rhône-Alpes – EdF.

[43] C. Breuils, « Modélisation d'un changement de comportement de maintenance à partir de données doublement censurées», rapport de DEA de statistique, université Denis Diderot, 2000.

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

[45] J. Diebolt, V. Durbec, C. Trottier, « Régularisation de distributions pour une meilleure adéquation extrême», 1999, rapport final de contrat Inria Rhône-Alpes – EdF.

[46] C. Goutte, C. Lavergne, Y. Vernaz, « Analyse de sensibilite du code ACTIGE », 2000, rapport de contrat Inria Rhône-Alpes – CEA Cadarache.

[47] C. Goutte, C. Lavergne, Y. Vernaz, «Application de methodes discriminantes pour scenarios d'accidents graves», 2000, rapport de contrat Inria Rhône-Alpes – CEA Cadarache.

[48] C. Lavergne, Y. Vernaz, «Indice de sensibilité», 2000, rapport de contrat Inria Rhône-Alpes – CEA Cadarache.

[49] D. Migault, *Modélisation statistique et fusion de caractéristiques de chrominance et d'orientation : Application à l'indexation de base d'images*, Mémoire, ENSERG 2ème année, INPG, 2000.

[50] R. Neugebauer, « Analyse statistique d'images IRM pour la détection de cancers du sein », rapport de DEA de Biostatistique, université de Montpellier 1, 2000.