INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet IS2

Inférence statistique pour l'industrie et la santé

Grenoble

THÈME 4A

Table des matières

1 Composition de l'équipe

Responsable scientifique

Gilles Celeux [DR Inria]

Personnel Inria

Anne Guérin-Dugué [CR Inria depuis le 1/10/99, détachée de l'INPG]

Anatoli Iouditski [CR Inria jusqu'au 30/9/99]

Florence Forbes [CR Inria]

Christian Lavergne [professeur, université Paul Valéry, Montpellier en délégation à l'Inria jusqu'au 31/8/99]

Personnel des établissements partenaires

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

Chercheurs post-doctorants

Christophe Biernacki [boursier Inria jusqu'au 30/9/99, maître de conférences, université de Besançon]

Sophie Lambert-Lacroix [ATER université Joseph Fourier, Grenoble 1 jusqu'au 30/9/99, maître de conférences, université Joseph Fourier]

Catherine Trottier [boursière Inria jusqu'au 30/9/99, maître de conférences, université de Lille 3]

Chercheurs doctorants

Henri Bertholon [enseignant CNAM]

Yann Vernaz [boursier Inria]

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

Cécile Delhumeau [CHU de Grenoble]

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

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

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

Stagiaires longue durée

Franck Corset [ENSAI, Rennes]

Collaborateurs extérieurs

Christine Cans [médecin, association RHEOPS]

Stéphane Chrétien [projet NUMOPT]

Jean Diebolt [DR CNRS LMC-SMS]

Jérôme Fauconnier [praticien hospitalier, CHU de Grenoble]

Anatoli Iouditski [professeur, université Joseph Fourier, Grenoble 1, depuis le 1/10/99]

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 Modèles à structure cachée

Participants : Christophe Biernacki, Isabel Brito, Gilles Celeux, Jean Diebolt, Jean-Baptiste Durand, Florence Forbes, Nathalie Peyrard.

Mots clés : données manquantes, mélange de lois, algorithme EM, algorithmes stochastiques, 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[MT97].

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

[MT97] G. J. McLachlan, K. T., *The EM algorithm and extensions*, John Wiley, New York, 1997.

adoptée dépend beaucoup du but poursuivi. Par exemple, dans un contexte de classification, l'objectif est surtout de restaurer les catégories manquantes z_i, alors que dans un contexte d'estimation de densités, il est plutôt d'estimer le paramètre θ. Cela étant, une approche répandue consiste à se placer dans un cadre bayésien non informatif et à chercher le modèle m qui maximise la vraisemblance intégrée[RL97]

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

$\pi(\theta \mid m)$ 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

[RL97] K. Roeder, W. L., «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.

nous noterons z_i (resp. y_i) la valeur de l'image z (resp. y) au pixel i et plus généralement z_A (resp. y_A) la restriction de z (resp. y) à un sous-ensemble A de pixels.

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

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

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

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

[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.

$z^{(0)}$, à l'itération $t+1$, un pixel i est choisi et sa valeur est mise à jour en lui donnant la valeur qui maximise $f(z_i \mid z_{S\setminus\{i\}}, y, \theta, \beta)$.

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

Algorithmes non supervisés Les méthodes indiquées ci-dessus supposent les paramètres θ et β connus. En pratique, ces paramètres doivent être estimés à partir des informations disponibles, ce qui peut présenter certaines difficultés dans le cas des modèles markoviens. Lorsque l'on dispose de données pour lesquelles on connaît à la fois les observations y et la vraie image z, on peut envisager d'estimer les paramètres β et θ lors d'une phase d'apprentissage. Très souvent, de telles données ne sont pas disponibles. Il arrive également que la phase d'apprentissage demande l'intervention d'un opérateur humain dans des situations où une automatisation du système est souhaitée. Ainsi, la recherche d'algorithmes non supervisés est-elle d'un grand 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 [13] un algorithme fournissant une segmentation z sans donner une estimation précise de β, ce qui permet d'éviter des calculs coûteux.

3.2 Modèles linéaires généralisés et hétéroscédasticité

Participants : Christian Lavergne, Catherine Trottier, Yann Vernaz.

Mots clés : Modèles linéaires généralisés, hétéroscédasticité, structure exponentielle, modèles à effets aléatoires, modèles 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 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 Identification non paramétrique et systèmes adaptatifs

Participants : Anatoli Iouditski, Sophie Lambert-Lacroix.

Mots clés : identification, systèmes adaptatifs, filtrage adaptatif, modélisation « black-box ».

Résumé : *On se donne une suite d'observations $(Y_k)_{0<k\leq N}$ de loi caractérisée par un paramètre inconnu $\theta \in \Theta$. À la différence des problèmes paramétriques, nous supposons que le paramètre à estimer est une fonction (un vecteur de dimension infinie) et l'ensemble Θ est une classe fonctionnelle. L'approche que nous adoptons est alors celle de l'estimation non paramétrique (fonctionnelle). Il s'agit plus précisement d'approximer le paramètre θ par un vecteur de grande dimension $\bar{\theta}$ et d'utiliser des méthodes statistiques spécifiques pour obtenir une bonne estimation $\widehat{\theta}_N$ de $\bar{\theta}$ à partir des observations disponibles. Les questions qui se posent sont*

alors : lorsque la taille N de l'échantillon $(Y_k)_{0<k\leq N}$ tend vers l'infini, 1) comment construire la bonne approximation $\bar{\theta}$ de dimension finie, 2) l'estimation $\widehat{\theta}_N$ converge-t-elle vers θ, et si oui, à quelle vitesse ? 3) Notre estimateur, est-il optimal, c.-à-d. pouvons-nous démontrer il n'existe pas d'autre estimateur θ^*_N de θ qui converge plus vite, etc.

3.3.1 Estimation non paramétrique

La question qui nous intéresse à présent est la suivante : comment prédire les sorties d'un système dynamique donné, quand on dispose de très peu d'informations relatives à la structure des relations entrées-sorties (on ne suppose pas que le système est linéaire). Dans le contexte de traitement du signal ce problème est nouveau; il est connu sous le nom de *modélisation « black-box »*. Par ailleurs, depuis de longues années les statisticiens étudient ce problème, connu sous le nom de « régression non paramétrique ». Sous sa forme la plus traditionnelle, le problème d'estimation non paramétrique se pose ainsi : étant donné N observations $y_i,\ x_i,\ i = 1, ..., N$, liées par la relation

$$y_i = f(x_i) + e_i,$$

estimer la fonction inconnue $f(x)$. Les x_i et e_i sont des variables aléatoires i.i.d. (indépendantes et identiquement distribuées). Nous mesurons l'erreur de l'estimation $\widehat{f}_N$ par le risque quadratique

$$\int (\widehat{f}_N(x) - f(x))^2 P(dx),$$

ou $P(\cdot)$ est la distribution de x_N. Il existe de nombreux algorithmes d'estimation de f, qui donnent des estimées $\hat{f}_n$, telles que sous certaines conditions de régularité de la fonction f on puisse majorer l'erreur d'estimation $\hat{f}_n - f$, i.e. telles que les vitesses de convergence des algorithmes classiques d'estimation non paramétrique dépendent étroitement de la régularité de la fonction à estimer. Or, dans certains problèmes de traitement du signal, le problème de débruitage de parole en donne un exemple, les signaux à restaurer sont loin de satisfaire des conditions de régularité. Dans l'exemple du traitement de parole ces sont des signaux qui oscillent rapidement, modulés par des fonctions régulières.

Quelquefois les régresseurs x_i appartiennent à un espace de grande dimension : dans ce cas l'approche classique d'estimation fonctionnelle est inopérante car les observations x_i sont trop espacées pour permettre une approximation efficace de la fonction inconnue. Des hypothèses spécifiques sur la structure de f permettent de baisser sa dimension « efficace» et de construire des estimateurs précis.

Une autre classe de problèmes intéressants est formée par les besoins de la théorie d'identification non linéaire pour un système dynamique du type

$$y_i = f(y_{i-1}, u_i) + e_i.$$

Dans cette situation, la loi de $x_i = (y_{i-1}, x_i)$ dépend elle-même de f inconnue, de plus, les (x_i) sont dépendants, ce qui peut ralentir sensiblement l'estimation (en particulier si la variance de e_i est petite).

4 Domaines d'applications

4.1 Fiabilité industrielle

Participants : Henri Bertholon, Christophe Biernacki, Gilles Celeux, Franck Corset, Jean Diebolt, Christian Lavergne, Catherine Trottier, 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 auquels 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'occurence 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 une 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, Jérome Fauconnier, Christian Lavergne, Claudine Robert.

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 médecins au CHU de Grenoble du projet et des 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.

5 Logiciels

5.1 Boîte à outils MATLAB de modélisation non linéaire

Participant : Anatoli Iouditski.

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

En coopération avec Lennart Ljung et Peter Lidskog de l'université de Linköping, Qinghua Zhang et Bernard Delyon de l'Irisa, Rennes, nous préparons, depuis l'automne 1996, une boîte à outils Matlab. Cette boîte à outils est conçue comme une extension de la boîte à outils System Identification (SI-Toolbox) de Lennart Ljung, qui servira à la modélisation de systèmes dynamiques non linéaires. Les techniques utilisées sont les algorithmes adaptatifs d'estimation non paramétrique, les réseaux de neurones et les réseaux d'ondelettes. Les modèles proposés sont pour l'essentiel de type auto-régressif non linéaire avec quelques extensions spécifiques pour lesquelles on dispose de bons algorithmes. La boîte à outils sera distribuée par Mathworks en 2000.

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 utilisation intensive d'algorithmes non itératifs d'estimation non paramétrique basés sur le triage adaptatif des estimées, *algorithmes d'arbre*, développé 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.

Etant complétement 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. Pour ces méthodes voir articles [Ja95],[DI97]. Outre les services d'identification proprement

[Ja95] A. JUDITSKY, AL, « Nonlinear Black-Box Modelling in System Identification », *Automatica 31*, 12, 1995, p. 1725–1750.

[DI97] B. DELYON, A. IOUDITSKI, « On minimax prediction for nonparametric autogressive models », *Publication interne IRISA*, 1121, 1997.

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, Gérard Govaert, Van Mô Dang.

Les mélanges multivariés gaussiens constituent un modèle de référence en analyse discriminante et en classification [MB88]. 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[CD85] 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[22]. 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 soi-même ses stratégies ou de s'en remettre à des choix par défaut. À terme, nous comptons le transcrire en Scilab. Deux évolutions essentielles sont prévues : l'extension à d'autres types de mélange, comme Bernoulli pour les données binaires ou Student (déjà intégré dans EMMIX) et l'extension vers des modèles de chaînes de Markov cachées.

5.3 Le logiciel EXTREMES

Participants : Myriam Garrido, Jean Diebolt, Catherine Trottier.

Dans le cadre de notre collaboration avec le groupe «Retour d'expérience» de la DER-EDF, nous avons programmé un logiciel interactif en Matlab, interne à EDF, et intitulé EXTREMES. Ce logiciel permet de réaliser toute la procédure du test ET (cf. section 7.3), et l'alternative basée sur une approche bayésienne (cf. section 5.3). Les quatre procédures proposées sont

- Un test d'adéquation classique,
- Un test d'adéquation de la loi exponentielle aux excès,

[MB88] G. McLachlan, K. E. Basford, *Mixture Models - Inference and Applications to Clustering*, Marcel Dekker, New York, 1988.

[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.

[CD85] G. Celeux, J. Diebolt, «The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorit hm for the mixture problem», *Computational Statistics Quarterly 2*, 1985, p. 73–92.

- 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 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 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 construit à partir de la quasi-vraisemblance au sens de Hutton-Nelson ; la convergence et les propriétés asymptotiques de ces estimateurs sont établies. Les avantages de l'algorithme sont sa simplicité de mise en œuvre, sa rapidité et sa stabilité numérique. La

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

construction d'une fonction quasi-score permet également d'établir les tests classiques d'hypothèses : tests de quasi-Wald, quasi-score et rapport de quasi-vraisemblance.

On s'intéresse aussi à l'estimation non paramétrique du modèle (3) par la méthode des polynômes locaux. Le point sensible est le choix de la fenêtre du noyau de convolution. Nous proposons une procédure adaptative du choix de fenêtre [LV97]. L'application de cet algorithme au modèle (3) permet d'obtenir une estimation optimale des fonctions $m(.)$ et $\sigma(.)$ en un point donné. Puis nous proposons ensuite d'appliquer une méthode itérative par polynômes locaux. Les simulations numériques effectuées sont encourageantes. Par contre l'étude théorique reste à faire [6], [47].

6.1.2 Modèles linéaires généralisés à effets aléatoires

Participants : Christian Lavergne, Catherine Trottier.

Les modèles linéaires généralisés mixtes (GL2M) ou à effets aléatoires sont la combinaison naturelle des modèles linéaires généralisés (GLM) et des modèles linéaires mixtes (L2M). Dans ces modèles, sous une hypothèse gaussienne de distribution des effets aléatoires ξ, la vraisemblance basée sur la distribution marginale du vecteur à expliquer Y n'est pas en général explicitement calculable. Diverses approximations peuvent être appliquées, basées soit sur une approche conditionnelle soit sur une approche marginale. Nous avons proposé dans un premier temps une méthode qui consiste en une maximisation de la distribution jointe de (Y, ξ) avant de procéder à l'estimation des paramètres. Ceci équivaut à une linéarisation conditionnelle du modèle [46], [16]. Une démarche marginale qui repose sur l'approximation des deux premiers moments marginaux de Y puis sur l'utilisation de la quasi-vraisemblance constitue la deuxième approche [18]. Nous introduisons ensuite une notion d'hétérogénéité dans les GL2M qui est modélisée en attribuant à chaque environnement un paramètre de variance différent pour ces effets. Nous proposons une méthode d'estimation combinant à la fois la technique de linéarisation de la démarche conditionnelle précédente et l'utilisation de l'algorithme EM.

6.1.3 Modèles linéaires généralisés auto-régressifs et conditionnellement hétéroscédastiques

Participant : Christian Lavergne.

Dans ce travail, nous nous proposons d'étudier des modèles auto-régressifs pour des séries chronologiques dont la loi de probabilité sous-jacente appartient à la structure exponentielle. Deux mémoires de DEA ont été proposés sur ce thème. Irwan Susanto (université J. Fourier) a étudié les particularités de certains de ces modèles régits par la loi de Poisson. En particulier les caractéristiques d'hétéroscédasticité conditionnelle et de queues de distribution lourdes sont établies dans le cas d'un modèle GLM-ARCH poisson d'ordre un. Sébastien Cosme (université de Montpellier) a étudié sur des simulations quatre modèles auto-régressifs pour des séries de fréquences. Cette étude a porté sur la qualité des estimations et des statistiques de tests.

[LV97] O. Lepski, S. V., « Optimal pointwise adaptative methos in nonparametric estimation », *Annals of Statistics 6*, 1997, p. 2512–2546.

6.2 Estimation non paramétrique

6.2.1 Commande adaptative

Participant : Anatoli Iouditski.

Une partie importante de la théorie de l'estimation statistique concerne l'établissement des limites intrinsèques de performances des algorithmes, et, par conséquent, fournissent une caractérisation pertinente du problème d'estimation en question ; ils donnent ainsi une échelle absolue d'optimalité pour toutes les techniques qui sont proposées pour la résolution de ce problème. Par rapport aux problèmes classiques d'estimation stochastique, le problème de la commande adaptative est assez singulier : il possède un degré de liberté supplémentaire qui est la commande.

Nous avons continué, en collaboration avec des chercheurs de l'IPU (Institute for Control Science) de Moscou, l'étude des algorithmes adaptatifs de commande pour des systèmes dynamiques non linéaires. Des nouveaux algorithmes de commande ont été proposés et leur efficacité a été établie ([17], [34]).

6.2.2 Estimation de paramètres macroscopiques

Participant : Anatoli Iouditski.

Dans un bon nombre de problèmes de modélisation non paramétrique appliqués à la finance la question qui intéresse le chercheur est d'établir une estimation de paramétres macroscopiques ou spatiaux de modèle inconnu. Un exemple classique de ce type est fourni par le problème d'estimation de l'indice spatial dans le modèle de type « single-indice » ou multi-indice. Nous proposons ([15]) une nouvelle méthode d'estimation du coefficient d'indice dans un modèle « single-indice », qui est basée sur des améliorations itératives de l'estimateur de la dérivée moyenne. L'estimateur qui en résulte est $\sqrt{n}$–consistant (n étant la taille de l'échantillon). Le travail de généralisation de cette méthode d'estimation très prometteuse dans le cadre de modèles multi-indice est en cours.

6.2.3 Intervalles de confiance pour des algorithmes adaptatifs

Participants : Anatoli Iouditski, Sophie Lambert-Lacroix.

Nous nous sommes intéressés à la construction des intervalles de confiance pour des estimateurs fonctionnels adaptatifs. En collaboration avec Oleg Lepski de l'université Aix-Marseille I, nous avons établi le cadre théorique précis d'estimation de la norme d'erreur. En outre, nous proposons des algorithmes adaptatifs d'estimation non paramétrique par ondelettes avec des intervalles de confiance associés. On démontre l'efficacité de ces algorithmes pour une grande variété de classes fonctionnelles ([33], [45]).

6.3 Modèles à structure cachée

6.3.1 Inférence bayésienne pour les mélanges

Participants : Gilles Celeux, Merrilee Hurn[1], Christian Robert[2].

Dans un contexte bayésien, nous nous sommes intéresser aux problèmes d'exploitation et d'interprétation de la loi a posteriori d'un modèle de mélange [41]. Intrinsèquement, la loi a posteriori d'un modèle de mélange présente $k!$ modes. Les méthodes classiques de Monte-Carlo par chaînes de Markov ont en général de grandes difficultés pour restituer ces modes très séparés : l'échantillonneur de Monte-Carlo reste dans le voisinage d'un mode sans parvenir à visiter les autres modes d'égales importances. Nous montrons que l'exploration de ces modes peut être imposée à l'échantillonneur de Monte-Carlo en utilisant des transitions de refroidissement (*simulated tempering*). Mais dans un cadre non informatif, comme la loi a priori traite les composants du mélange de manière indifférenciée, la loi a posteriori est symétrique et les estimateurs classiques comme les moyennes a posteriori sont inutilisables. Ils le restent d'ailleurs avec des lois a priori informatives qui distinguent les composants entre eux. Nous proposons différentes solutions pour mener à bien l'inférence bayésienne à partir de lois a posteriori invariantes par permutation des indices des composants du mélange. L'une utilise une technique de classification, l'autre se fonde sur la minimisation de fonctions de perte appropriées. Un à-côté important de cette étude est la mise en évidence de l'efficacité des techniques de refroidissements pour accélérer la convergence des algorithmes de Monte-Carlo par chaînes de Markov.

6.3.2 Accélération de l'algorithme EM pour les mélanges

Participants : Gilles Celeux, Stéphane Chrétien, Florence Forbes, Abdallah Mkhadri[3].

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 [40] 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 l'algorithme SAGE [FH94] qui vise également à accélérer l'algorithme EM.

6.3.3 Approximation du champ moyen et segmentation d'images

Participants : Gilles Celeux, Florence Forbes, Nathalie Peyrard.

1. university of Bath
2. Crest-Insee
3. université de Marrakech

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

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 son utilisation, dans le cadre de la segmentation markovienne d'images, comme outil algorithmique pour éviter les calculs coûteux inhérents aux modéles 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 approximation peut être encore 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. Dans le cas des modèles markoviens, deux difficultés se présentent : le calcul de la fonction de partition et celui des probabilités marginales a posteriori. Des approximations ont été proposées pour traiter ces étapes. 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 le principe du champ moyen, qui contient et généralise la procédure proposée par Zhang. Cette famille d'algorithmes inclut également d'autres procédures connues (telle que l'algorithme NEM d'estimation de mélanges sous contraintes spatiales), que nous présentons sous une structure commune.

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

Participants : Florence Forbes, Chris Fraley [4], Nathalie Peyrard, Adrian Raftery [4].

Cette étude se place dans le cadre d'une collaboration entre Toshiba MRI inc. à San Francisco, l'université de Washington à Seattle et le projet IS2 de l'Inria Rhône-Alpes. Le contexte est celui de l'Imagerie par Résonance Magnétique (IRM) pour la détection de tumeurs cancéreuses du sein. L'utilisation de cette technique est fondée sur la grande réactivité des cellules cancéreuses après administration d'un agent de contraste (Gadolinium). Les experts de Toshiba ont sélectionné 8 variables qui résument la forme du signal reçu pour chaque pixel et qui sont susceptibles d'aider à la caractérisation des tumeurs. Cependant, l'idéal pour les radiologues est d'avoir une seule image différenciant clairement les tumeurs des tissus sains. Nous avons réalisé des premiers travaux qui utilisent des méthodes statistiques de segmentation, spatiales et multivariées, pour synthétiser l'information disponible en une seule image. Une première étape utilise une méthode de classification de données multivariées EMCLUST[FR99]. Elle ne prend pas en compte l'aspect spatial et débouche souvent sur des classifications comportant des régions trop inhomogènes. Une deuxième étape consiste donc à améliorer ces premières

4. Statistics Department, University of Washington, Seattle

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

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

[FR99] C. Fraley, A. E. Raftery, «MCLUST: Software for model-based cluster analysis», *Journal of Classification*, 1999, à paraître.

classifications à l'aide de méthodes spatiales. Dans notre étude, nous avons utilisé un algorithme de *morphologie bayésienne* ([13]) et des filtres morphologiques. Les premiers résultats obtenus ont fait l'objet d'un rapport ([43]).

6.3.5 Formation de la feuille de papier

Participants : Gilles Celeux, Florence Forbes, Nathalie Peyrard.

Cette étude est l'objet d'une collaboration avec des chercheurs de l'école de papeterie de Grenoble (Jean-Francis Bloch) et de l'université de Beira Interior au Portugal (Jacques Silvy et Ana Paula Costa). Il s'agit d'étudier la *formation* de la feuille de papier dans le but de rendre compte des différents types de papiers et d'étudier la dépendance aux conditions de fonctionnement des machines.

Un premier type d'approches consiste à travailler directement sur les images en niveaux de gris sans considération du processus de formation. Il s'agit essentiellement de décrire la répartition des niveaux de gris observés en calculant des mesures caractéristiques (moyenne, variance, entropie, etc.). Un second type d'approches consiste à partir d'un modèle du processus de formation pour en déduire un modèle de la feuille. Un exemple est le modèle de floculation de Farnood *et al.*[FDL95] qui s'avère décevant. Nous proposons de nous placer entre ces deux types d'approches. Il s'agit de donner un modèle des images en niveaux de gris sans nécessairement que celui-ci résulte d'une connaissance précise du phénomène physique sous-jacent.

L'aspect nuageux et aléatoire des papiers nous a conduit naturellement à examiner des modèles de type champ de Markov. Nous avons effectué un prétraitement des images, de type segmentation, pour isoler et identifier les *flocs* (agrégats de fibres, plus ou moins cohérents). Leurs caractéristiques (régularité, taille, association, orientation, etc.) semblent jouer un rôle important pour les propriétés physiques et les qualités du papier. Nous avons obtenu des segmentations des images en niveaux de gris en quatre classes à l'aide d'un algorithme de type ICM (morphologie bayésienne de [13]). La deuxième étape a consisté à étudier les caractéristiques de ces flocs et leurs associations. Nous nous sommes intéressés à la distribution de leurs surfaces, périmètres et à leur répartition dans la feuille de papier. Ces premiers résultats ont été exposés et les programmes permettant de les obtenir installés à l'université du Beira Interior au Portugal.

D'autre part, une approche concurrente d'analyse de textures[DSP99] est envisagée en collaboration avec X. Descombes de l'Inria Sophia-Antipolis et P. Perez de l'Irisa.

6.3.6 Classification de suites finies par des chaînes de Markov cachées

Participants : Gilles Celeux, Jean-Baptiste Durand, Florence Forbes.

Dans le cadre de son stage de DEA, Jean-Baptiste Durand a eu à traiter au sein de l'équipe

[FDL95] R. R. Farnood, C. Dodson, S. R. Loewen, « Modeling Flocculation. Part I: Random Disk Model », *Journal of Pulp Paper Science 21*, 10, 1995, p. 348–356.

[DSP99] X. Descombes, M. Sigelle, F. Préteux, « GMRF Parameter Estimation in a non-stationary Framework by a Renormalization Technique: Application to Remote Sensing Imaging », *IEEE Transaction on image processing 8*, 4, 1999, p. 490–503.

PRIMA de l'Imag un problème de reconnaissance de gestes par une chaîne de Markov cachée. D'un point de vue méthodologique, il s'agit d'un problème de classification qui ne diffère des problèmes d'analyse discriminante que par le fait que les données à classer sont des suites finies. Le but est de construire une règle de décision à partir d'un échantillon d'apprentissage étiqueté. L'approche choisie consiste à modéliser les suites à classer par des chaînes de Markov cachées. Un modèle de Markov caché est associé à chaque classe a priori. Ses paramètres sont estimés à partir des données d'apprentissage de la classe associée, en utilisant l'algorithme EM. L'un des problème à résoudre concerne le choix du nombre d'états pour chaque modèle de Markov. La solution proposée se base sur le critère BIC. Cette approche a été mise en œuvre avec succès dans le cadre de la reconnaissance de tracés de lettres effectués du doigt devant une caméra([52]). La méthode du choix du nombre d'états cachés est comparée à une méthode heuristique basée sur le nombre d'arcs et de segments du tracé.

6.3.7 Indexation d'images

Participant : Christophe Biernacki.

Dans le cadre d'une année post-doctorale dans le projet MOVI en liaison avec IS2, Christophe Biernacki a appliqué avec succès le modèle de mélange à des problèmes d'indexation d'images [21]. Cette recherche est décrite dans le rapport d'activité du projet MOVI.

6.4 Algorithmes stochastiques

6.4.1 Contrôle de la convergence

Participants : Didier Chauveau[5], Jean Diebolt.

Dans le contexte actuel de l'utilisation intensive en statistique des méthodes de Monte-Carlo par chaînes de Markov (MCMC), il est essentiel de s'assurer que les résultats proposés sont valides et précis. nous avons mis au point une nouvelle méthode de contrôle de la convergence vers la stationnarité des chaînes de Markov engendrées par les MCMC, méthode qui permet la construction de « régions de confiance » pour les ensembles de quantités à évaluer. Notre procédure de contrôle utilise des chaînes simulées en parallèle, afin de déterminer au moyen d'approximations empiriques un nombre minimal d'itérations au-delà duquel l'utilisation du théorème de la limite centrale puisse être considérée comme justifiée, en vue de construire des intervalles de confiance pour les résultats fournis par l'algorithme. Ce travail a fait en 1999, sous une forme améliorée, l'objet de l'article [11]. Nous avons pu tester la procédure sur de nombreux exemples. L'accès au logiciel (Mathematica et C) peut se faire via http://www-math.univ-mlv.fr/~chauveau/cltc.html.

6.4.2 Algorithmes de recherche stochastiques

Participants : Olivier François[6], Christian Lavergne.

5. université de Marne-la-Vallée
6. LMC, Imag

Dans ce travail, nous proposons une méthodologie statistique destinée à aider l'utilisateur d'algorithmes évolutionnaires à configurer ces algorithmes. L'outil statistique permet de piloter de manière pertinente et efficace le plan de simulations inhérent à toute phase de réglage. La méthodologie présentée permet de prendre en compte efficacement les expériences croisées ou passées. En utilisant l'information contenue dans toutes les simulations, un modèle statistique permet de rationaliser les plans d'expérience. En particulier, pour de nombreux problèmes, il est possible de gérer un très petit nombre de simulations pour chaque combinaison de paramètres.

Les procédures d'estimation et de tests permettent d'identifier les effets de la variation des paramètres de l'algorithme sur la solution produite. La possibilité de constituer des classes de problèmes permet d'améliorer la qualité des estimations au même coût de simulation. Le modèle statistique utilisé dans ce travail est un modèle de type boîte noire. Il ne prend en compte aucune propriété géométrique du problème à minimiser. Il modélise la qualité des solutions produites par l'algorithme et non le fonctionnement de l'algorithme lui-même. Dans la plupart des problèmes réalistes, aucune propriété analytique du problème à résoudre n'est accessible. C'est en ce sens que la méthodologie statistique se justifie [44].

6.5 Choix de modèles en discrimination et classification automatique

6.5.1 Vraisemblance complétée intégrée

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

Le critère BIC, approximation de la vraisemblance intégrée, est sans doute le critère le plus populaire pour évaluer le nombre de classes d'un mélange. Pourtant, il présente un défaut pratique. Il s'avère peu robuste car trop sensible à la violation des hypothèses du modèle de mélange. Aussi, nous avons conçu un critère, dénommé ICL (*Integrated Completed Likelihood*), qui se présente comme une approximation de la vraisemblance intégrée des données complétées. Pour compléter les données manquantes, nous affectons les points observés aux classes inconnues par un opérateur du maximum a posteriori. Cette année nous avons construit une simplification du critère ICL qui ne nécessite pas de définir de loi a priori pour les paramètres du mélange. Il s'exprime maintenant comme une pénalisation du critère BIC par un terme d'entropie de la partition estimée par maximum de vraisemblance. L'ancienne et la nouvelle version sont très proches par leur expression et leur comportement. De nouvelles expérimentations viennent aussi renforcer l'intérêt du critère ICL.

6.5.2 Combinaison de modèles en analyse discriminante

Participants : Isabel Brito, Gilles Celeux, Ana Maria Sousa Ferreira[8].

Ce thème constitue le sujet des thèses que prépare 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

7. UTC, Compiègne
8. université de Lisbonne

que celles tirées d'une seule méthode. Isabel Brito avait montré l'intérêt des règles linéaires fondées sur la vraisemblance intégrée et des combinaisons hiérachiques de modèles [23]. Elle explore actuellement trois nouvelles directions de recherche. La première calcule les poids des modèles en minimisant l'erreur définie comme l'espérence de la différence quadratique entre la caractérisation observée et la caractérisation prédite par le modele combiné. Cette technique s'apparente au *Committee of methods* de Bishop[Bis95], pour la deuxième approche le schéma d'échantillonnage des données doit être celui du mélange. Alors, la combinaison de modèles peut ainsi être vu comme un modéle de mélange de mélanges et l'algorithme EM est utilisé dans l'estimation des poids, les composants étant fixes. La troisième approche complémentaire vise à éviter de combiner des modèles trop semblables par une sélection préalable utilisant des outils exploratoires classiques d'analyse des vecteurs d'affectation.

Dans le cadre qualitatif où travaille Ana Maria Sousa Ferreira, il est possible, en se dotant de lois a priori non informatives, de calculer sans approximation la vraisemblance intégrée du modèle multinomiale complet et du modèle d'indépendance conditionnelle pour combiner ces deux modèles de manière optimale [29]. Mais, actuellement, elle se heurte à des problèmes de résolution numériques pour des problèmes de grande taille. Par ailleurs, une approche de couplage hiérarchique de méthodes utilisant le coefficient d'affinité de Matushita a été envisagée [29].

6.6 Modèles de fiabilité industrielle

6.6.1 Un modèle de vieillissement

Participants : Henri Bertholon, Gilles Celeux.

Dans le cadre d'une convention d'étude et recherche avec le département « Sûreté de fonctionnement, Diagnostic, Maintenance» de EDF-DER, nous avons proposé un nouveau modèle de vieillissement apparenté à la loi de Weibull mais qui introduit un instant de début de vieillissement t_0 strictement positif. Nous avons étudié ce modèle pour des matériels réparables et non réparables. Nous avons construit des procédures d'estimation du maximum de vraisemblance, le paramètre de forme gouvernant le vieillissement étant fixé. Il nous est apparu en effet irréaliste de pouvoir estimer tous les paramètres dans un contexte industriel où les défaillances sont rares.

Par ailleurs, nous avons construit un test optimal de l'existence d'un vieillissement obéissant à notre modèle, lorsque le paramètre d'échelle de la loi exponentielle qui décrit l'absence de vieillissement est connue. Diverses expérimentations numériques sur des données simulées et réelles ont montré la cohérence de nos procédures et l'intérêt de notre modèle.

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

Participants : Gilles Celeux, Franck Corset, Christian Lavergne, Yann Vernaz.

[Bis95] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

Cette recherche traite de deux 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 défaillance 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.

Par ailleurs, nous avons eu à traiter, lors d'une nouvelle convention de recherche avec le département «Sûreté de fonctionnement, Diagnostic, Maintenance» de EDF-DER, des données de durée de vie de pompes de même type sur un matériel non réparable, mais où la modélisation par un processus de Poisson n'était pas envisageable car les dégradations n'étaient pas rares. Aussi, nous avons pu utiliser le maximum de vraisemblance pour estimer les paramètres d'une loi exponentielle et d'une loi de Weibull sans renconter les difficultés précédentes. Nous avons pu également, sous certaines conditions, donner les propriétés de normalité asymptotique de ces estimateurs et donc en déduire des intervalles de confiance. Par ailleurs une approche bayésienne a été considéré dans un cadre informatif (présence d'avis d'experts), non informatif et semi-informatif (on sait juste que le matériel ne peut que vieillir). Cette dernière méthode s'est révélée la plus intéressante. Enfin, nous avons considéré un procédure non paramétrique d'estimation de la fonction de survie (cf Turnbull[Tur74]) dont nous avons montré qu'elle peut être vu comme un algorithme EM, où les censures à gauche sont considérées comme les valeurs manquantes du modèle.

6.6.3 Modélisation et estimation de queues de distributions

Participants : Jean Diebolt, Myriam Garrido, Catherine Trottier.

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

[Tur74] B. Turnbull, «Non parametric estimation of a survivorship function with doubly censored data», *Journal of the American Statistical Association 69*, 1974, p. 1974.

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 (testant principalement la partie centrale de la distribution) aboutissent à des conclusions différentes de celles de ce test extrême. Ainsi, le présent contrat tente de répondre à la question : « Quel nouveau modèle peut-on proposer lorsqu'un modèle a été accepté par un test d'adéquation global et rejeté par un test extrême ? ».

Notre point de vue est de proposer une 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. Nous avons pour cela envisagé des solutions par recollement de la partie centrale et d'une approximation ET de la partie extrême, par modèles de mélanges gaussiens après une transformation normalisante. Enfin nous avons aussi adopté un point de vue bayésien sur cette régularisation en supposant une loi a priori sur l'un des paramètres de la loi du modèle. Les hyperparamètres peuvent être ajustés de différentes façons selon que cette loi a priori correspond ou non à la prise en compte d'un avis d'expert (à défaut elle permettra de réaliser un compromis entre l'estimation du maximum de vraisemblance du paramètre et son estimation ET). La loi régularisée proposée est alors la loi prédictive obtenue en sortie de l'inférence bayésienne en réalisant un mélange continu des lois de la famille du modèle selon la loi a posteriori du paramètre.

6.7 Statistique biomédicale

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

6.7.1 Évaluation de l'exhaustivité d'un registre de morbidité.

Participants : Christine Cans, Pascale Guillem, Christian Lavergne.

Notre étude a porté sur neuf générations d'enfants nés entre 1980 et 1988, domiciliés en Isère et porteurs d'au moins une déficience sévère, selon les critères d'inclusion du RHEOP (Registre des Handicaps de l'Enfant et Observatoire Périnatal). Ces enfants avaient été recrutés à partir de quatre sources de données différentes. L'exhaustivité du registre a été estimée, dans un premier temps par la méthode de capture-recapture à partir de deux sources dont nous avons montré l'indépendantes par la méthode de Wittes. Dans un second temps, l'analyse statistique s'est faite par des modèles log-linéaires. Leur intérêt réside dans l'absence de contraintes liées

[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.

au respect des conditions de validité nécessaires avant d'appliquer la méthode de capture-recapture, et dans la possibilité de prendre en compte non seulement les quatre sources, mais aussi des variables d'hétérogénéité comme par exemple le nombre de déficiences par enfant [9], [24].

6.7.2 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 très invalidante. Elle compromet l'autonomie de l'enfant et implique une prise en charge lourde. C'est l'une des maladies les plus commune chez les jeunes enfants. Sa prévalence est de 3/1 000 naissances vivantes, ce taux pouvant augmenter jusqu'à 80 / 1 000 dans les groupes à risques (bébés prématurés ou trop petits). L'aide nécessaire varie grandement en fonction du degré de dépendance de la personne.

Le projet européen a pour but de créer, à travers l'Europe, un réseau de registres de morbidité basé sur l'étude des enfants atteints de CP. Actuellement, nous constituons une base de données exploitable commune à 14 centres européens regroupant des caractéristiques médico-sociales de ces enfants. Cette base de données va nous permettre de fournir des estimations fiables du taux de prévalence de la CP en Europe. Cette base de données devrait aussi permettre d'identifier les facteurs de risque de cette maladie qui sont aujourd'hui, pour la plupart, non démontrés vu le faible nombre de cas disponibles dans chaque pays.

6.7.3 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. Il est intéressant de 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 des 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 atypiques et d'essayer de les classer selon leurs distributions. Nous avons modélisé ces distributions de quantiles de DS par un mélange de lois normales en utilisant le logiciel XEMGAUS. Une classification en 5 groupes ressort nettement de cette modélisation. Parmi ces groupes, trois ressortent comme mauvais (DS plus longues que la moyenne). Il s'agit maintenant d'analyser finement ces trois groupes et en particulier de voir si cette allongement des durées de séjour n'est pas dû à un recrutement particulier à Grenoble.

7 Contrats industriels (nationaux, européens et internationaux)

7.1 Contrat EDF : Évaluation de la constance du taux de défaillance

Participants : Henri Bertholon, Gilles Celeux.

Ce contrat de type CERD (Convention d'Études et de Recherche) avec le département « Surveillance, Diagnostic, Maintenance » de la DER-EDF concernait l'analyse de durées de vie et en particulier visait à proposer des méthodes pour une mise en évidence rapide d'un éventuel vieillissement. Nous avons ainsi effectué une revue bibliographique des principaux tests de vieillissement en distinguant bien les situations où les matériels sont réparables ou non. Cette étude nous a conduit à proposer le modèle décrit en 6.6.1 qui peut être vu comme un modèle de Weibull où le vieillissement est retardé ce qui correspond bien à une situation pratique courante.

7.2 Retour d'expérience de constituants de pompes

Participants : Gilles Celeux, Franck Corset.

Ce contrat de type CERD avec le département « Surveillance, Diagnostic, Maintenance» de 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 s'est faite à partir de nombreuses bases de données bien renseignées et où les dégradations sont nombreuses. Par contre, on ne connaît pas les instants de dégradation. Seules des données censurées à gauche et à droite sont disponibles. Nous avons utilisés des modèles exponentiel et de de Weibull estimés par maximum de vraisemblance ou par une approche bayésienne (cf. [50]).

7.3 Contrat EDF sur les queues de distribution de probabilité

Participants : Jean Diebolt, Myriam Garrido, Catherine Trottier.

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 de Jean Diebolt et Stéphane Girard sur ce même thème l'an dernier. 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. Dans le cadre d'un autre contrat Catherine Trottier, a proposé différentes approches lorsque aucun des modèles paramétriques standards utilisés pour le test extrême ne pouvait représenter à la fois le comportement global et le comportement extrême. La suite de ce travail va consister à élargir la plage des lois dont nous pouvons tester l'adéquation aux extrêmes, et à développer les alternatives proposées par Catherine Trottier, notamment celles basées sur une approche bayésienne.

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.

8 Actions régionales, nationales et internationales

8.1 Actions régionales

IS2 participe régulièrement au séminaire de statistique du LMC-SMS à Grenoble et Gilles Celeux est l'un des organisateurs. Dans ce cadre, plusieurs conférenciers ont été invités. De plus, cette année, J. Diebolt, A. Guérin-Dugué, C. Robert et Y. Vernaz ont exposé à ce séminaire.

A. Iouditski a été invité au séminaire de statistique de l'université Paris VI (Jussieu), de l'ENS Paris et de l'université de Aix-Marseille.

Gilles Celeux a été invité au séminaire de statistique de Vannes.

X. Descombes et P. Perez ont effectué un séjour d'une semaine au sein du projet IS2, dans le cadre d'un projet de collaboration entre les projets VISTA (UR de Rennes), ARIANA (UR de Sophia-Antipolis) et IS2, sur le thème des modèles probabilistes spatiaux.

G. Celeux, F. Forbes et N. Peyrard collaborent avec des chercheurs de l'école de papeterie de Grenoble et l'université de Beira Interior au Portugal sur le thème de la formation du papier. F. Forbes a effectué dans ce cadre un séjour d'une semaine à l'université de Beira Interior.

J. Diebolt a été rapporteur de la thèse d'Isabelle Deltour, « Modélisation bayésienne de données avec erreurs de mesure et de données manquantes dans un contexte épidémiologique» : thèse (université Paris XI) préparée sous la direction de Sylvia Richardson (Inserm, Unité 170, Villejuif).

8.2 Actions nationales

Forte implication de IS2 dans l'organisation des XXXIèmes journées de statistique de la SfdS99, Grenoble. C. Lavergne était vice-président du comité d'organisation, G. Celeux membre du comité scientifique et C. Robert a animé une table ronde sur l'enseignement de la statistique et organisé la session de bio-statistique.

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. Il s'est rendu à Lisbonne en décembre pour travailler à l'avancement de la thèse d'A. M. Sousa Ferreira et participer à un atelier sur les problèmes de validation en classification automatique. Par ailleurs, il poursuit la direction de la thèse d'I. Brito, assistante au département d'économie de l'université de Lisbonne, sur la comparaison de méthodes d'analyse discriminante.

A. Iouditski a été chercheur invité, pour un séjour de deux semaines, à l'institut de Weierstrass (WIAS) à Berlin en juin 1999.

J. Diebolt a poursuivi la direction de la thèse de Jacques Zuber, École Polytechnique Fédérale de Lausanne, sur les tests d'adéquation de modèles de régression linéaire généralisée (GLM). Cette thèse a été soutenue le 4 juin 1999.

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. À cette occasion, elles ont participé au groupe de travail «Model-Based Clustering and Applications» organisé par A. Raftery.

8.5 Accueil de chercheurs étrangers

R. Liptser (université de Tel-Aviv) et A. Nemirovski (Technion, Haifa) ont passé un mois à l'Inria Rhône-Alpes pour poursuivre leur collaboration avec A. Iouditski. A. Nasin (université de Moscou) a été invité pour un séjour de 2 mois. Enfin, A. Mkhadri (université de Marrakech) a été invité un mois.

9 Diffusion de résultats

9.1 Animation de la communauté scientifique

A. Iouditsky est responsable d'un projet CNRS sur le filtrage adaptatif, en collaboration avec B. Delyon (université de Rennes) et E. Moulines (ENST).

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.

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).

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 le cours de « tests d'adéquation non paramétriques » dans le DEA de mathématiques appliquées de Grenoble. Jean Diebolt a également assuré un cours de 12 heures à l'Ensae, Paris, en mars–avril 1999, aux élèves de deuxième année sur la théorie des tests. Il s'agit d'un cours de niveau intermédiaire entre maîtrise et DEA. Il a aussi participé à la préparation de la nouvelle épreuve orale (calcul scientifique et modélisation) de l'agrégation de mathématiques.

F. Forbes assure des TDs de probabilités en première année à l'Ensimag.

9.3 Participation à des colloques, séminaires, invitations

N. Peyrard a participé à la *Second European conference on highly structured stochastic systems*, Pavia, 14-18 septembre 1999.

I. Brito a participé au *IXth International Symposium on Applied Stochastic Models and Data Analysis*, Lisbonne, 14-17 juin 1999.

G. Celeux a participé aux deuxièmes ateliers de recherche d'EDF en octobre 1999.

F. Forbes a été invitée au Séminaire de Statistique Européen (Semstat) du 15 au 20 mars 1999 à Eindhoven, Pays-bas.

Les membres d'IS2 ont participé aux XXXIèmes journées de statistique de la SfdS, à Grenoble en mai 1999.

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] 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.

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

Thèses et habilitations à diriger des recherches

[6] 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, Grenoble, 1999.

Articles et chapitres de livre

[7] C. Biernacki, G. Celeux, G. Govaert, «An Improvement of the NEC Criterion for Assessing the Number of Clusters in a Mixture Model», *Pattern Recognition Letters 20*, 1999, p. 267–272.

[8] C. Biernacki, G. Govaert, « Choosing Models in Model-based Clustering and Discriminant Analysis», *Journal of Statistical Computation and Simulation*, à paraître.

[9] C. Cans, P. Guillem, C. Lavergne, « Estimation of Morbidity Register Completeness Ascertainment cases, example from the RHEOP », *Revue d'Épidémiologie et de Santé Publique*, à paraître.

[10] G. Celeux, M. Persoz, J. Ngatchou-Wandji, F. Perrot, « Using Markov Chain Monte Carlo methods to solve full Bayesian modeling of PWR vessel flaw distributions », *Reliability Engineering and System Safety 66*, 1999, p. 243–252.

[11] D. Chauveau, J. Diebolt, «An automated stopping rule for MCMC convergence assessment», *Computational Statistics 14*, 3, 1999, p. 419–442.

[12] J. Diebolt, J. Zuber, «Goodness-of-fit tests for nonlinear heteroscedastic regression models», *Statistics and Probability Letters 42*, 1999, p. 53–60.

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

[14] S. Girard, J. Diebolt, « Consistency of the ET method and smooth variations», *Notes aux Comptes Rendus de l'Académie des Sciences de Paris, Série I*, 1999, à paraître.

[15] M. Hristache, A. Juditsky, V. Spokoiny, «Direct Estimation of the Index Coefficients in a Single-index Model», *Annals of Statistics*, à paraître.

[16] C. Lavergne, C. Trottier, « Sur l'estimation dans les modèles linéaires généralisés à effets aléatoires», *Revue de Statistique Appliquée 1.*

[17] A. Nazin, A. Juditsky, « Attainable information bounds in adaptive control of nonlinear stochastic systems under nonparametric uncertainty», *Avtomat. Telemekh.*, 3, 1999, p. 180–196.

[18] C. Trottier, « A quasi-marginal approach in generalized linear mixed models», *Statistics*, à paraître.

Communications à des congrès, colloques, etc.

[19] K. Aubert, P. Bryla, G. Celeux, C. Lavergne, Y. Vernaz, « A degradation model integrating prior kinetics parameters and Qualitative in-service inspection data», *in : ESREL'99, European Safety and Reliability Conference*, p. 743–748, TUM Munich-Garching, Allemagne, 13-17 septembre, 1999.

[20] C. Biernacki, F. Beninel, V. Bretagnolle, « Model on a Population and Prediction on Another One: Generalized Allocation Rule», *in : TMBM'99, Theory and Mathematics in Biology and Medicine*, Amsterdam, Pays-Bas, 29 juin-3 juillet, 1999.

[21] C. Biernacki, R. Mohr, «Indexation et appariement d'images par modèle de mélange gaussien des couleurs», *in : GRETSI'99, 17ème colloque GRETSI sur le traitement du signal et des images*, Vannes, France, 13-17 septembre, 1999.

[22] C. Biernacki, «Le logiciel XEMGAUS pour la classification et la discrimination gaussienne», *in : XXXIèmes journées de statistique*, Grenoble, France, 17-21 mai 1999.

[23] I. Brito, G. Celeux, «Combining EDDA's Models», *in : Applied Stochastic Models and Data Analysis*, p. 136–141, Lisbonne, Portugal, 14-17 juin, 1999.

[24] C. Cans, P. Guillem, C. Lavergne, «L'exhaustivité d'un registre et le modèle log-linéair e», *in : XXXIèmes journées de statistique*, Grenoble, France, 17-21 mai 1999.

[25] G. Celeux, F. Forbes, N. Peyrard, « Approximation du champ moyen et segmentation d'images», *in : XXXIèmes journées de statistique*, Grenoble, France, 17-21 mai 1999.

[26] G. Celeux, F. Forbes, N. Peyrard, «Mean field approximation principle and Markov random fields model-based image segmentation.», *in : Model-Based Clustering, Spatial Data Analysis, and Bayesian Model Selection Workshop*, Seattle - USA, 26 -29 juillet, 1999.

[27] G. Celeux, F. Forbes, N. Peyrard, «Mean field approximation principle and Markov random fields model-based image segmentation», *in : Second European Conference on Highly Structured Stochastic Systems*, Pavia - Italy, 14 - 18 septembre, 1999.

[28] D. Chauveau, J. Diebolt, «CLT convergence control», *in : Atelier "MCMC–orics" du TMR HSSS, CREST–Ensae*, Malakoff, France, décembre 1998.

[29] A. Ferreira, G. Celeux, H. Bacelar-Nicolau, «Combining Models in Discriminant Analysis and Hierarchical Coupling Approach», *in : Applied Stochastic Models and Data Analysis*, p. 159–164, Lisbonne, Portugal, 14-17 juin, 1999.

[30] S. Girard, J. Diebolt, «Consistance de la méthode ET et variations régulières», *in : XXXIèmes journées de statistique*, Grenoble, France, 17-21 mai 1999.

[31] S. Girard, J. Diebolt, « Convergence de l'estimation des quantiles extrêmes », *in : Colloque de la Société Mathématique Tunisienne*, Tabarka, Tunisie, mars 1999.

[32] R. Hammoud, R. Mohr, C. Biernacki, « Robustification des signatures de couleurs par modélisation de leur variabilité intra-plan vidéo», *in : GRETSI'99, 17ème colloque GRETSI sur le traitement du signal et des images*, Vannes, France, 13-17 septembre, 1999.

[33] A. Juditsky, S. Lambert-Lacroix, «Intervalles de confiance pour les estimateurs adaptatifs sur les classes de Besov », *in : XXXIèmes journées de statistique*, Grenoble, France, 17-21 mai 1999.

[34] A. Nazin, A. Juditsky, « Information Lower Bounds and Optimal Algorithms for Nonlinear Adaptive Control Problem», *in : Proceedings of 6th Saint Petersburg Symposium on Adaptive Systems Theory dedicated to the memory of Ya.Z. Tsypkin, SPAS'99*, 1, p. 113–117, St.-Petersburg, Russia, 7-9 septembre 1999.

[35] B. Villain, B. Vérité, C. Biernacki, G. Celeux, «A Practical Approach of Expert Elicitation for Bayesian Reliability Analysis of Ageing», *in : ESREL'99, European Safety and Reliability Conference*, p. 707–712, TUM Munich-Garching, Allemagne, 13-17 septembre, 1999.

[36] J. Zuber, J. Diebolt, «Un test chi-carré d'adéquation de modèles paramétriques en régression», *in : XXXIèmes journées de statistique*, Grenoble, France, 17-21 mai 1999.

[37] J. ZUBER, J. DIEBOLT, « A goodness-of-fit test for nonlinear models based on nonparametric techniques», *in : Inverse Problems in Statistics, Mathematisches Forschungsinstitut*, Oberwolfach, Allemagne, janvier 1999.

[38] J. ZUBER, J. DIEBOLT, « A goodness-of-fit test for nonlinear models based on nonparametric techniques », *in : Séminaire de printemps du 3ème cycle romand de statistique et probabilités appliquées*, Villars, Suisse, mars 1999.

Rapports de recherche et publications internes

[39] C. BIERNACKI, R. MOHR, «Indexation et appariement d'images par modèle de mélange gaussien des couleurs», *rapport de recherche n° 3600*, Inria Rhône-Alpes, Grenoble, 1999.

[40] G. CELEUX, F. FORBES, A. MKHADRI, S. CHRÉTIEN, « A Component-Wise EM algorithm for Mixtures», *rapport de recherche n° 3746*, Inria Rhône-Alpes, Grenoble, 1999.

[41] G. CELEUX, M. HURN, C. ROBERT, « Computational and Inferential Difficulties with Mixture Posterior Distributions», *rapport de recherche n° 3627*, Inria Rhône-Alpes, Grenoble, 1999.

[42] S. CHRÉTIEN, A. HERO, « Kullback Proximal Point Algorithm for Maximum Likelihood Estimation», *rapport de recherche n° 3756*, Inria Rhône-Alpes, Grenoble, 1999.

[43] F. FORBES, C. FRALEY, N. PEYRARD, A. E. RAFTERY, « Spatial Statistical analysis of breast Magnetic Resonance Image via Model-based Clustering, Morphology ang Markov random fields», *rapport de recherche*, For Toshiba MRI inc, Août 1999.

[44] O. FRANÇOIS, C. LAVERGNE, « Plans d'expériences pour l'évaluation d'algorithmes évolutionnaires et la constitution de classes de référence», *rapport de recherche n° 3601*, Inria Rhône-Alpes, Grenoble, 1999.

[45] A. JUDITSKY, S. LAMBERT-LACROIX, « Confidence intervals for adaptive regression estimation on the Besov spaces», *rapport de recherche n° 3643*, Inria Rhône-Alpes, Grenoble, France, 1999.

[46] C. LAVERGNE, C. TROTTIER, « Sur l'estimation dans les modèles linéaires généralisés à effets aléatoires», *rapport de recherche n° 3630*, Inria Rhône-Alpes, Grenoble, 1999.

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

Divers

[48] H. BERTHOLON, G. CELEUX, A.LANNOY, «Évaluation de la constance du taux de défaillance», Rapport final de convention de recherche Inria-EDF, 1999.

[49] H. BERTHOLON, G. CELEUX, A.LANNOY, « Proposition de correction de la norme IEC 56/33/537/FDIS concernant le test d'hypothèse du taux de défaillance constant », Lettre au secrétaire de la CEI en charge de la norme 56/537/FDIS, 1999.

[50] G. CELEUX, F. CORSET, M.-A. GARNERO, «Analyse du retour d'expérience de constituants de pompe», Rapport final de convention de recherche Inria-EDF, 1999.

[51] J. DIEBOLT, V. DURBEC, C. TROTTIER, « Régularisation de distributions pour une meilleure adéquation extrême», Rapport final de convention de recherche Inria-EDF, 1999.

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