INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Projet IS2

Inférence statistique pour l'industrie et la santé

Rhône-Alpes

THÈME 4A

Table des matières

1. Composition de l'équipe

Responsable scientifique
Gilles Celeux [DR Inria]

Personnel Inria
Florence Forbes [CR Inria]
Paulo Gonçalvès [CR Inria]

Personnel des établissements partenaires
Christian Lavergne [professeur, université Paul Valéry, Montpellier]
Claudine Robert [professeur, université Joseph Fourier, Grenoble 1]

Chercheurs post-doctorants/Ingénieurs
Edwige Allain [boursière Inria depuis le 01/12/02]
Gérard Boudjema [boursier Inria]
Emilie Lebarbier [boursière Inria depuis le 01/10/02]
Jérome Ecarnot [ingénieur expert depuis le 01/03/02]
Grégory Noulin [ingénieur expert depuis le 01/10/02]

Chercheurs doctorants
Guillaume Bouchard [boursier Inria]
Franck Corset [boursier Inria]
Jean-Baptiste Durand [boursier MESR]
Julien Jacques [boursier Inria]
Christophe Lenoir [Inserm U572]
Olivier Martin [boursier MESR]
Carine Vera [Cirad]
Matthieu Vignes [boursier AC depuis 01/09/02]

Collaborateurs extérieurs
Henri Bertholon [enseignant CNAM, Paris]
Jean Diebolt [DR CNRS université de Marne-la-Vallée]
Myriam Garrido [ATER Grenoble 2]
Stéphane Girard [maître de conférences, université Joseph Fourier, Grenoble 1]
Anatoli Iouditski [professeur, université Joseph Fourier, Grenoble 1]

Assistante de projet
Françoise de Coninck

2. Présentation et objectifs généraux

Le projet IS2 effectue des recherches en modélisation statistique. Plus spécifiquement, nous nous intéressons à la modélisation, à l'identification des modèles obtenus et à leur validation pour des systèmes ou des situations complexes pouvant intervenir dans le domaine industriel ou biomédical.
IS2 s'intéresse essentiellement aux modèles, dits à structure de données incomplètes, où intrinsèquement une partie de l'information nécessaire à l'identification du phénomène étudié est manquante. Ces modèles sont courants (durées de vie censurées, modèles hétéroscédastiques[1], images dégradées, ...) et puissants (modèles à structure cachée, ...). Ils apparaissent dans de nombreux problèmes statistiques qui se posent en milieu biomédical et en milieu industriel. Ces modèles à observation partielle sont difficiles à estimer, de par leur nature intrinsèque et aussi parce qu'ils concernent eux-mêmes des systèmes complexes (montages industriels compliqués, existence d'une structure de dépendance temporelle ou spatiale, nombreuses variables en jeu,...).

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

De ce fait, ces modèles sont en général faiblement identifiables en ce sens que, au vu des observations effectivement recueillies, plusieurs jeux différents de paramètres peuvent apparaître également bons. Cela se traduit par une multiplicité des *extrema* locaux des fonctions de contraste utilisées pour procéder à l'identification (vraisemblance, probabilité a posteriori,...). Ainsi, ces modèles requièrent une grande rigueur conceptuelle et méthodologique, le recours raisonné à un principe de parcimonie (retenir le modèle le moins complexe pour une qualité d'ajustement acceptable), et l'utilisation d'outils algorithmiques sophistiqués.

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

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

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

3. Fondements scientifiques

3.1. Modèles à structure cachée

Participants : Guillaume Bouchard, Gilles Celeux, Jean-Baptiste Durand, Florence Forbes, Paulo Gonçalvès, Olivier Martin, Matthieu Vignes.

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

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

3.1.1. Généralités

Le projet IS2 s'intéresse à des modèles statistiques paramétriques, θ étant le paramètre à estimer, où les données complètes $x = x_1, \cdots, x_n$ se décomposent de manière naturelle en données observées $y = y_1, \cdots, y_n$ et en données manquantes $z = z_1, ..., z_n$. 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 3.1.2.1).

3.1.1.1. 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[71].

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.

3.1.1.2. Choix de modèles

Un point important pour les modèles à structure cachée est le choix de la complexité du modèle et en particulier le choix du nombre K de catégories de la variable cachée. Dans ce domaine, très ouvert, de nombreuses approches sont en compétition et la stratégie adoptée dépend beaucoup du but poursuivi. Par exemple, dans un contexte de classification, l'objectif est surtout de restaurer les catégories manquantes z_i, alors que dans un contexte d'estimation de densités, il est plutôt d'estimer le paramètre θ. Cela étant, une approche répandue consiste à se placer dans un cadre bayésien non informatif et à chercher le modèle m qui maximise la vraisemblance intégrée[74]

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

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

3.1.1.3. 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[72],

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 soulèvent de nombreuses questions de modélisation et d'inférence statistique et n'ont cessé de gagner de l'intérêt. En particulier, le choix de modèles appropriés et l'estimation des paramètres associés aux modèles utilisés sont des questions essentielles pour aller vers une automatisation des algorithmes et tirer tout le profit de la richesse des modèles stochastiques. Ces problèmes, abondamment traités, restent cependant ouverts. En effet, un effort d'ordre méthodologique (recherche d'estimateurs précis et robustes) et d'ordre algorithmique (réduction des temps de calcul) reste à faire.

3.1.2.1. Segmentation et restauration d'image

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

Une approche possible, bien fondée statistiquement, est l'analyse d'image dite bayésienne. Elle fournit des solutions élégantes et a connu des développements considérables depuis des premiers travaux tels que ceux de D. et S. Geman[65] ou Besag[60]. 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

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

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[65] 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[65] 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[60]. La convergence n'est toutefois garantie que vers un maximum local de (1) et l'algorithme peut être très sensible aux conditions initiales. À partir d'une image initiale $z^{(0)}$, à l'itération $t+1$, un pixel i est choisi et sa valeur est mise à jour en lui donnant la valeur qui maximise $f(z_i \mid z_{S\setminus\{i\}}, y, \theta, \beta)$.

3.1.2.2. 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.

3.1.2.3. 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 [7] un algorithme fournissant une segmentation z sans donner une estimation précise de β, ce qui permet d'éviter des calculs coûteux.

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

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

Le modèle statistique ainsi défini sur les coefficients d'ondelette est un modèle à structure cachée pour lequel existent des algorithmes de calcul et de maximisation de la vraisemblance comparables à l'algorithme avant-arrière pour les chaînes de Markov cachées [20].

Ainsi, si l'on privilégie l'axe temporel, on s'attache à modéliser la dépendance statistique de l'état d'un système conditionnellement à son passé et relativement à une échelle de temps donnée. À l'inverse, si on

privilégie l'axe des échelles, on cherche à caractériser les interactions entre plusieurs échelles de temps. On peut ainsi envisager de repérer grâce à ces modèles, des comportements en loi d'échelle (auto-similarité globale ou locale, longue dépendance), ou, ce qui nous intéresse davantage, des transitions dans cette dynamique d'échelle (processus multi-échelle, scalings non stationnaires...).

3.2. Modèles linéaires généralisés et hétéroscédasticité

Participants : Christian Lavergne, Christophe Lenoir, Carine Véra.

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

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.

3.2.1. 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.

3.2.2. 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 $y = (y_1, \cdots, y_n)$ s'écrit :

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

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

ii) d'un prédicteur linéaire $\eta = (\eta_i)_{i=1,\cdots,n}$ 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 $\eta = X\beta$.

3.2.3. 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 $i = 0, ..., p$ 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.

3.2.4. Les modèles linéaires généralisés mixtes

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

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

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

3.2.5. Les modèles glm-arch

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

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

les valeurs initiales η_1,..., η_q étant fixées, de sorte que la vraisemblance conditionnelle de y sachant le passé soit celle d'un GLM avec comme prédicteur linéaire η_t.

3.3. Estimation de lois d'échelle par ondelettes

Participant : Paulo Gonçalvès.

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

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

Soit $(t, x(t))$ la trajectoire d'un processus aléatoire. La régularité hölderienne locale de $x(t)$ est définie par [6][11]

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

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

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

où $\dim_{\mathcal{H}}\{E\}$ désigne la dimension de Hausdorff de l'ensemble E. En pratique cette définition se heurte à plusieurs difficultés. D'une part il n'est pas possible d'accéder en chaque point t de la trajectoire de x à la régularité hölderienne $\alpha(t)$. D'autre part, on ne sait pas calculer efficacement la dimension de Hausdorff associée à chacun des sous-ensembles de points de même régularité $\alpha(t) = \alpha$. L'*analyse multifractale* permet de contourner ces difficultés en proposant des alternatives au spectre de Hausdroff telles que le spectre de Legendre, ces alternatives pouvant même dans certains cas désignés par *formalisme multifractal*, conduire à des équivalences strictes entre les spectres.

Une version très répandue du spectre de Legendre fait usage des coefficients d'ondelette issus de la décomposition du signal x

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

dont on étudie les moments d'ordre $q \in \Re$ estimés empiriquement par

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

Les lois d'échelle qui structurent le processus x se traduisent par un comportement en loi de puissance de ces moments à travers les échelles, selon

$$S^n(q) \equiv 2^{n\tau(q)}.$$

Le *spectre de Legendre* correspond alors simplement à la transformée de Legendre de la fonction $\tau(q)$:

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

En toute généralité on a la relation $d(\alpha) \leq f(\alpha)$, entre spectre de Hausdorff et spectre de Legendre, l'égalité pouvant être atteinte pour les processus vérifiant le formalisme multifractal.

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

4. Domaines d'application

4.1. Fiabilité industrielle

Participants : Henri Bertholon, Gérard Boudjema, Gilles Celeux, Franck Corset, Jean Diebolt, Julien Jacques, Christian Lavergne, Myriam Garrido, Stéphane Girard.

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 R&D. Les problèmes auxquels nous sommes confrontés relèvent de l'analyse de durées de vie de systèmes non réparables pouvant être sujets à vieillissement, l'étude de la cinétique de dégradation de systèmes passifs (tuyaux par exemple) et la modélisation statististique de modes de défaillance prenant en compte l'avis d'experts. Les données dont nous disposons pour ces études viennent du retour d'expérience associé aux

opérations de maintenance préventive. Elles sont alors de nature quantitative. Sinon il s'agit d'avis d'experts le plus souvent qualitatifs.

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

Glossaire

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

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

Glossaire

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

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

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

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

4.2. Statistique biomédicale

Participants : Gilles Celeux, Paulo Gonçalvès, Olivier Martin, Christian Lavergne, Christophe Lenoir, Claudine Robert.

Notre deuxième domaine d'intervention, moins développé, concerne les applications biomédicales. Les problèmes que nous considérons s'orientent selon trois axes distincts :

- l'analyse de données hospitalières ou la détermination de facteurs de risque de maladies. Ils se concrétisent dans le cadre d'actions avec les collaborateurs extérieurs du projet, médecins au CHU de Grenoble (Jérome Fauconnier), et membres du laboratoire TIMC de l'Imag. Nous sommes amenés à mettre en œuvre des modèles assez variés de type modèle linéaire et des techniques d'analyse multidimensionnelle des données (arbres d'induction, analyses factorielles) ;
- l'analyse de données issues de puces à ADN (ou biopuces). Ce domaine connait un développement important en raison des problèmes statistiques qui y sont liés et des résultats attendus en génomique fonctionnelle. À partir de mesures donnant le niveau d'expression de plusieurs milliers de gènes, il s'agit de déterminer leurs implications dans des processus biologiques. Notre travail actuel s'est axé sur les problèmes de normalisation et de recherche de gènes différentiellement exprimés. Dans la suite, nous souhaitons en tirer des éléments pour aborder la classification des profils d'expression ;
- l'analyse du rythme cardiaque chez la souris. En utilisant des modèles linéaires généralisés à effet mixte, nous voulons caractériser les mécanismes de régulation engendrés par le système nerveux autonome sur l'activité cardiaque. En particulier, nous cherchons à identifier la nature, vagale ou sympathique, du tonus dominant chez ce mammifère.

5. Logiciels

5.1. Le logiciel mixmod

Participants : Gilles Celeux, Grégory Noulin.

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

MIXMOD a été écrit en C++ et des interface Scilab et Matlab ont été réalisés. Il est diffusé en *open source* et téléchargeable à l'adresse web suivante : http://www-math.univ-fcomte.fr/MIXMOD/index.htm. Il sera prochainement diffusé sur le cdrom des logiciels libres distribués par l'Inria.

Dans le cadre d'un ODL INRIA, nous avons reçu le renfort d'un ingénieur expert, Grégory Noulin, pour améliorer MIXMOD dans les directions suivantes.

- Améliorer les performances d'exécution de MIXMOD afin d'être compétitifs vis-à-vis des techniques exploratoires traditionnelles. Au-delà d'une optimisation du code par son examen critique, le travail consiste aussi à concevoir et à réaliser une version de MIXMOD spécifique pour le traitement de très gros jeux de données.
- Réalisation d'un site web de qualité comportant les rubriques suivantes : known bugs, FAQ, Bug report, how to, register, newsgroup...
- Actuellement, MIXMOD propose des interfaces graphiques pour Matlab et Scilab. Il s'agit de réaliser une fonction Scilab à part entière, ce qui permettra entre autres de profiter des fonctions graphiques de Scilab.
- Au plan des modèles proposés, il s'agit d'étendre le logiciel aux cas où les observations sont décrites par des variables binaires ou qualitatives par le traitement des mélanges de Bernoulli et multinomiaux multivariés. C'est un domaine où nous avons développé une panoplie de modèles d'une richesse et d'une nature analogue à ce que nous avons fait dans le cas gaussien En particulier MIXMOD pourra ainsi traiter conjointement des variables quantitatives et qualitatives sous certaines hypothèses.
- Dans le souci d'extension de MIXMOD, on proposera des outils pour une mise à jour régulière du logiciel, réaliser une documentation complète et soignée, assurer la plus grande transparence pour le développement (fichiers sources commentés, diagrammes de conception UML publiés, ...).

La nouvelle version de MIXMOD, sortie à la fin de cette année, contient des entrées par mots-clés, la possiblité de pondérer les individus, une installation automatique du logiciel sous Windows et une première version complète de la page web.

5.2. Le projet SEL

Participant : Claudine Robert.

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

Le portail web (http://www.inrialpes.fr/is2/) de statistique en ligne à l'usage des enseignants en mathématiques du secondaire a été diffusé à 30 000 exemplaires (tous les professeurs de mathématiques des lycées) et a déjà été demandé par divers pays. Un document d'accompagnement pour les nouveaux programmes des classes de seconde, première, terminale (programmes consultables à l'adresse http://www.education.gouv.fr/bo) ainsi que diverses animations sur des parties des programmes ont été ajoutés. Les textes de programme doivent être assez courts, précis tout en respectant la liberté de chaque enseignant. Ils sont associés à des documents d'accompagnement (environ 150 pages élaborées en 2001-2002 pour les classes de terminales et plus de 300 pages depuis 1999) qui illustrent les objectifs du programme, précisent des choix possibles, suggèrent des thèmes de travail. Ces documents peuvent être téléchargés à partir de l'adresse suivante http://www.eduscol.education.fr/D0015.

En plus des documents papiers, le groupe a conçu en 2001-2002 un CD contenant, outre les programmes des matières scientifiques et les documents d'accompagnement, le logiciel SEL d'autoformation à la statistique et diverses appliquettes, notamment en java, illustrant les nouveautés des programmes.

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

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

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

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

5.3. Le logiciel Extremes

Participants : Jean Diebolt, Jérôme Ecarnot, Myriam Garrido, Stéphane Girard.

Dans le cadre d'une collaboration avec EDF, Myriam Garrido a construit une maquette de logiciel sous Matlab implémentant les outils développés durant sa thèse. Nous reprenons en partie ses travaux pour programmer un logiciel en C++ dont les sources seront libres. Diverses interfaces utilisateurs sont aussi prévues pour en faciliter le maniement. Ce logiciel reprend les fonctionnalités du précédent auxquelles s'ajoutent un certain nombre d'outils de base dans le monde des extrêmes ainsi qu'une extension du test ET au test GPD. Pour l'instant, le logiciel permet les actions suivantes :

- Simulations de variables aléatoires de lois classiques.
- Graphique des densités, fonctions de répartition, fonctions de survie, et fonctions quantiles associées.
- Estimation des paramètres de lois classiques.
- Estimation de la densité (méthode de noyau et histogramme).
- Test d'Anderson-Darling et de Cramer-Von Mises.
- Test d'exponentialité des excès.
- Test ET (versions asymptotique, bootstrap et bootstrap simplifié) .
- Régularisation bayésienne.

- Estimateurs classiques de l'indice des valeurs extrêmes (Hill, Hill généralisé, Moments pondérés d'Hosking et Wallis, Maximum de vraisemblance et Zipf).
- Estimation des quantiles extrêmes.

Le code C++ est actuellement interfacé avec le logiciel Matlab.
D'autre part, il nous reste à implémenter :

- Le test GPD.
- Une procédure d'estimation bayésienne des paramètres de la loi GPD.
- Le passage à un autre système d'exploitation (Unix).
- Diverses interfaces utilisateurs.
- L'aide du logiciel.

Ce logiciel devra être livré à EDF à la fin du mois de février 2003.

6. Résultats nouveaux

6.1. Modèles à structure cachée

6.1.1. Échantillonnage préférentiel pour les modèles à structure cachée

Participant : Gilles Celeux.

Cette recherche s'effectue en collaboration avec Jean-Michel Marin et Christian Robert du Ceremade (université Paris 9 Dauphine). L'inférence bayésienne s'est beaucoup développée ces dernières années grâce au développement des techniques de chaînes de Markov de Monte-Carlo comme l'échantillonnage de Gibbs. Cependant, passé l'enthousiasme des premiers succès, on se rend compte maintenant que ces techniques ne sont pas dépourvues de défauts et peuvent notamment converger trop lentement. Aussi, les techniques d'échantillonnage préférentiel constituent de plus en plus une approche concurrente. Nous nous sommes intéressés à l'emploi de ces techniques pour des modèles à structure de données manquantes où la simulation des lois a posteriori de l'inférence bayésienne est en général nécessaire mais reste souvent difficile. Notre approche est de proposer différentes techniques d'échantillonnage préférentiel qui tirent parti de l'existence de données manquantes pour les simuler suivant différentes stratégies. À l'heure actuelle, nous avons exploré différentes pistes qui se distinguent essentiellement suivant le fait que les données manquantes sont simulées suivant la loi a priori ou après avoir déterminé l'estimateur du maximum de vraisemblance. De plus, nous comparons actuellement les différentes possibilités introduites sur la base d'études de cas.

6.1.2. Algorithmes d'inférence pour les modèles de Markov cachés

Participant : Jean-Baptiste Durand.

L'estimation de paramètres dans les modèles graphiques à variables aléatoires cachées (ou *modèles de Markov cachés*) par l'algorithme EM et ses variantes conduit à calculer la loi conditionnelle d'états cachés sachant les données observées. Ces calculs sont infaisables par marginalisation directe de la probabilité des données complètes et sont réalisés soit de manière approchée, soit, lorsque cela est possible, en utilisant des algorithmes récursifs efficaces, analogues à des parcours du graphe d'indépendance conditionnelle.

Nous avons défini une famille de modèles de Markov cachés pour laquelle l'algorithme EM admet une implémentation efficace. Des algorithmes résolvant entièrement le problème du calcul de probabilités (*i.e.* le calcul de $P(\mathbf{X}|\mathbf{Y})$ pour n'importe quels processus aléatoires $\mathbf{X}$ et $\mathbf{Y}$ obtenus par restriction du graphe d'indépendance conditionnelle) ont été proposés. Ces algorithmes sont basés sur une récursion arrière-avant utilisant l'arbre de jonction. Les formules de récursion utilisent les paramètres du modèle de manière explicite et présentent le grand avantage d'avoir une interprétation probabiliste et d'être numériquement stables. Ces algorithmes permettent le traitement de modèles graphiques avec des variables aléatoires à valeurs continues. De plus, un algorithme du MAP (*Maximum A Posteriori*) a été proposé pour la restauration des états cachés.

Ceci permet d'obtenir une variante *à la Viterbi* de l'algorithme EM. Enfin, une solution pour l'implémentation des variantes stochastiques de EM, à savoir SEM et EM *à la Gibbs*, a été donnée. Le problème de la simulation des états cachés dans SEM est résolu de manière efficace par une utilisation parcimonieuse de l'algorithme arrière-avant.

6.1.3. Sélection de modèles de chaînes de Markov cachées

Participants : Gilles Celeux, Jean-Baptiste Durand.

Le modèle de chaînes de Markov cachées est fréquemment utilisé en reconnaissance statistique des formes, notamment en reconnaissance de parole ou de gestes. Comme pour les mélanges de lois, l'un des problèmes qui reste à résoudre concerne le choix du nombre d'états cachés.

Nous avons entrepris de l'attaquer en utilisant une évaluation de la déviance du modèle par des techniques de validation croisée généralisant le *leave-one-out*, comme la validation croisée *multifold* et l'apprentissage/test répété. Dans le principe, cela consiste à diviser plusieurs fois l'échantillon en deux parties de taille éventuellement inégale, puis à estimer les paramètres sur une partie et à calculer la vraisemblance du modèle sur l'autre. On obtient ainsi une vraisemblance moyenne qui sert de critère de sélection. Du fait de la dépendance markovienne, le découpage en deux parties n'est pas une opération anodine. Dans le cas où les deux parties sont tirées au hasard, cela nous a amené à adapter l'algorithme de BAUM-WELCH de calcul de l'estimateur du maximum de vraisemblance dans une chaîne de Markov cachée à observations manquantes. Dans le cas où la chaîne est divisée suivant la parité des indices, nous avons montré que les processus obtenus sont encore des chaînes de Markov cachées, ce qui permet d'utiliser l'algorithme de BAUM-WELCH pour l'estimation des paramètres [38] et donc donne lieu à une procédure de sélection dite PI (pour Paire et Impaire) simple et rapide.

Les critères de sélection de modèles concurrents à la validation croisée sont essentiellement le critère BIC et le critère de vraisemblance marginale pénalisée VMP de Gassiat [64]. Les expérimentations menées sur des données simulées ont montré que même pour une séparation faible des composants du mélange, les critères BIC, VMP et la procédure PI, se fondant sur un partionnement alternatif et équilibré de l'échantillon, sélectionnent très fréquemment le modèle ayant généré les données (*modèle générateur*). Les critères de validation croisée *multifold* ont une tendance plus forte à surestimer le nombre d'états cachés. Dans le cas de mélanges très peu séparés, le nombre de données nécessaire aux critères BIC et PI pour déterminer le modèle générateur est nettement inférieur à celui nécessité par VMP.

Par ailleurs, Jean-Baptiste Durand en collaboration avec Olivier Gaudoin (LMC) a mené des applications des modèles de Markov cachés pour analyser la fiabilité de logiciels. À cette occasion, ces critères de sélection de modèles ont été expérimentés sur des données réelles. Cette étude a mis en évidence leur comportement similaire sur la plupart des jeux de données : les modèles sélectionnés sont parcimonieux. Cependant, les critères VMP et BIC, dans de rares cas, favorisent des modèles sans doute trop simples, ce qui appauvrit l'interprétation des temps inter-défaillances, en termes de recherche de zones homogènes dans le processus des corrections. A contrario, les méthodes de demi-échantillonnage de type PI notamment sélectionnent des modèles légèrement plus complexes, mais d'interprétation plus riche.

6.1.4. Mélange de régressions

Participant : Guillaume Bouchard.

Ce travail concernait les modèles à structure cachée dans un cadre de régression. Ces modèles sont utiles lorsque des hétérogénéités dans les données peuvent être observées. Des modèles appelés mélanges d'experts permettent de traiter ce cas. Le travail a consisté à étudier un cas particulier de ces modèles où l'on suppose que les variables explicatives suivent une loi gaussienne. Plusieurs modèles différents peuvent alors être envisagés en fonction des contraintes sur les paramètres des composants des mélanges et conduire à des structure de matrices de variance sophistiqués mais pouvant s'avérer utiles [29].

6.1.5. Modèles de chaînes de Markov cachées pour le suivi de contours

Participant : Gilles Celeux.

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

Nous avons développé un modèle de chaîne de Markov cachée et un algorithme pour son estimation par le maximum de vraisemblance pour un problème de suivi de dynamique avec comme application en vue le suivi des mouvements de lèvres d'une personne parlant ou chantant (*lip tracking*). L'intérêt et la difficulté viennent de ce que le signal observé est lui-même régi par un modèle autorégressif. Les paramètres de ce processus autorégressif varient en fonction des états cachés qui eux sont la réalisation d'une chaîne de Markov homogène. Un algorithme du maximum de vraisemblance de type EM a été écrit et programmé. Des expérimentations sur des données simulées ont montré son intérêt. Une application à un problème réel de suivi de lèvres est prévu. Cette recherche s'effectue dans le cadre de la collaboration INRIA/ICCTI (Portugal).

6.1.6. Analyse statistique d'Images à Résonance Magnétique (irm) pour la détection et l'identification de tumeurs

Participant : Florence Forbes.

Travail en collaboration avec Chris Fraley et Adrian Raftery (Statistics Department, University of Washington, Seattle), Bradley Wyman, Insightful inc. Seattle.

L'étude en collaboration avec l'université de Washington [63] suit son cours. Nous avons poursuivi le travail en cherchant à valider la méthode sur un plus grand nombre de patients ce qui a essentiellement demandé un important travail d'implémentation et d'adaptation des données. Nous avons du nous familiariser avec divers formats d'images et programmes de conversion, notamment pour passer du format dicom utilisé en IRM a des formats d'images plus classiques utilisables par nos programmes.

6.1.7. Modèles à structure de covariance pour la classification de données spatiales : application à des données issues d'IRM

Participants : Florence Forbes, Christian Lavergne.

Une approche classique pour la classification de données spatiales est de considérer un modèle de champ de Markov caché. Les dépendances spatiales entre pixels proches sont prises en compte au niveau des classes (variables cachées) à l'aide d'un modèle de champ de Markov discret. Les variables observées ont chacune une loi (par exemple gaussienne) dont les paramètres dépendent de la variable cachée (classe du pixel) associée. Une hypothèse de travail importante et simplificatrice est ici que les variables observées sont conditionnellement indépendantes sachant les données cachées.

Il s'agit d'étudier une autre approche en considérant un modèle global qui précise "directement" la structure de covariance des données observées. Une différence importante avec l'approche précédente est que l'hypothèse d'indépendance conditionnelle n'est plus faite. En revanche, dans un premier temps on supposera les variables cachées (les classes) indépendantes. Nous nous intéressons à la mise en œuvre d'un tel modèle, notamment sur les données IRM [63] qui ont la particularité de présenter des dépendances spatiales et temporelles.

Notre point de départ est le modèle CAR (conditional autoregressive) (voir par exemple Besag 1974 et plus précisément une extension spatio-temporelle proposée par Pettitt, Weir et Hart 2001 [73]. Il s'agit d'adapter ce modèle à notre objectif de classification, notamment en introduisant des variables cachées. Pour ce qui est de la mise en œuvre sur des données réelles, une étude préliminaire [55] montre que l'on se heurte rapidement au problème de l'inversion d'une matrice creuse de grande taille. Nous étudions donc entre autres l'utilisation d'algorithmes de type algorithme de Lanczos pour le calcul des valeurs propres dans ce cadre et le recours à une formulation variationnelle d'un tel problème.

6.2. Méthodes pour le choix de modèles

6.2.1. Critères de vraisemblance pénalisée

Participants : Gilles Celeux, Jean-Baptiste Durand, Florence Forbes, Christian Lavergne, Emilie Lebarbier.

Travail en collaboration avec Gabriela Ciuperca, université Lyon 1 et Frédérique Letué, université Grenoble 2.

Nous nous intéressons à la genèse et aux propriétés des critères de type vraisemblance pénalisée. Par exemple, une propriété importante est la consistance du critère, *i.e.* son aptitude à sélectionner le *meilleur* modèle lorsque le nombre de données augmente. Pour ce qui est de la sélection du nombre de classes d'un mélange E. Gassiat[64] a montré qu'un critère de vraisemblance marginale pénalisée était consistant sous l'hypothèse que la loi dont sont issues les données appartient à l'ensemble des modèles que l'on cherche à comparer. En pratique, cela est rarement le cas et il est important de pouvoir s'affranchir de cette hypothèse et de voir dans quelle mesure cela limite ou remet en cause le bien fondé de certains critères. Certains, tels le critère AIC (Akaike Information Criterion), ne sont clairement justifiés que sous cette hypothèse. Dans ce cadre nous avons mis en place un groupe de travail sur la sélection de modèle auquel participent également Guillaume Bouchard et Matthieu Vignes.

6.2.2. Critère DIC pour la sélection de modèles à structure cachée

Participants : Gilles Celeux, Florence Forbes.

Travail en collaboration avec Mike Titterington, université de Glasgow, Écosse et Christian Robert, CEREMADE, université Paris Dauphine.

Nous nous intéressons au critère DIC (Deviance Information Criterion) de sélection de modèle introduit dans [75]. Son utilisation dans le cadre des modèles à structure cachée n'est pas immédiate et nous proposons différentes possibilités d'extension du critère en fonction de la manière dont les données manquantes sont prises en compte. Il s'agit d'un critère introduit dans un cadre bayésien comportant une partie adaptation aux données, sous forme d'un terme de déviance, et une partie pénalisant la complexité du modèle. Nous considérons essentiellement trois approches. Dans le cadre bayésien, l'approche la plus naturelle est de considérer les données manquantes comme des paramètres supplémentaires et d'utiliser la vraisemblance conditionnelle sachant les données manquantes. Deux autres approches sont possibles : selon que les données manquantes sont explicitement prises en compte ou non, on considèrera la vraisemblance des données complètes ou celle des données observées. En pratique, ces trois approches peuvent conduire à des résultats assez différents. Pour commencer nous étudions leur comportement dans le cadre de la sélection de mélanges gaussiens.

6.2.3. Approximation variationnelle de critères de sélection

Participants : Gilles Celeux, Florence Forbes.

Travail en collaboration avec Mike Titterington, université de Glasgow.

Dans [22], nous avons proposé un critère basé sur le critère BIC et l'approximation du champ moyen. Nous avons constaté de bonnes performances pratiques mais nous ne disposons pas de résultats sur d'éventuelles propriétés théoriques, telles que la consistance. Un objectif est donc la définition et l'étude de nouveaux critères définis à partir de critères existants (dans un premier temps) et d'approximations de type champ moyen (variationnelles). Nous nous sommes par exemple intéressés à l'utilisation d'approximations variationnelles pour le critère DIC dans le cadre des modèles à structure cachée.

6.2.4. Mélange de lois normales sphériques en analyse discriminante

Participants : Guillaume Bouchard, Gilles Celeux.

L'idée de cette méthode destinée à proposer une alternative flexible à la discrimination linéaire lorsque cette méthode ne donne pas de résultats satisfaisants consite à modéliser la distribution de probabilité de chaque groupe par un mélange de lois normales de matrices variances proportionnelles à la matrice identité. De la sorte, nous proposons une méthode paramétrique parcimonieuse susceptible de produire des frontières de décision fortement non linéaires. L'un des obstacles que nous avons à surmonter est le choix du nombre de composants des mélanges. Des essais avec le critère BIC sont assez encourageants, mais il est souhaitable de définir une stratégie prenant en compte l'objectif de discrimination.

Par ailleurs, nous avons établi des liens algébriques entre l'approche de l'analyse discriminante fondée sur l'estimation des densités par groupe et l'approche fondée sur l'estimation directe des probabilités conditionnelles d'appartenance aux groupes. Ce rapprochement formel de ces deux grandes approches de l'analyse

discriminante nous ouvre des pistes pour le contrôle et l'amélioration de l'analyse discrimante par des mélanges de lois normales sphériques.

6.3. Modèles de fiabilité industrielle

6.3.1. Un modèle de vieillissement

Participants : Henri Bertholon, Gilles Celeux.

Suite à la thèse de Henri Bertholon [59] soutenue l'an dernier, nous nous sommes intéressé à un modèle de vieillissement qui nous semble d'un grand intérêt car il met en concurrence un décès dû à un vieillissement avec un décès accidentel. De la sorte, nous définissons une loi de probabilité qui s'écrit comme le minimum d'une loi exponentielle et d'une loi de Weibull. Notre loi dépend donc de trois paramètres qui sont les paramètres d'échelle des deux lois exponentielle et de Weibull et du paramètre de forme de la loi de Weibull. Nous avons étudié les caractéristiques théoriques de cette loi et proposé son estimation par le maximum de vraisemblance et par inférence bayésienne en incluant le cas courant où les données sont censurées à droite. L'estimation du maximum de vraisemblance se fait par l'algorithme EM et donne de bons résultats pour des échantillons d'assez petite taille. L'estimation bayésienne, utile pour des échantillons de très faible taille, se fait par un algorithme d'échantillonnage préférentiel qui tire parti du fait que ce modèle est un modèle à données manquantes pour tirer au hasard les données manquantes afin d'obtenir une distribution instrumentale réaliste. Les qualités de cet algorithme sont en cours d'évaluation. Ce travail d'estimation a fait l'objet du stage de DEA de biostatistique de Carole Langlois (Université de Montpellier) [54].

6.3.2. Un modèle de choc

Participant : Gilles Celeux.

Cette recherche se fait en collaboration avec Andrei Rodionov de l'IPSN. Il s'agit de caractériser la durée de vie d'un matériel soumis à des chocs lors de sollicitations et pouvant aussi être défaillant pour des raisons accidentelles indépendantes des chocs. Nous avons conçu un modèle à risques concurrents où les défaillances dues aux chocs obéissent à une loi Gamma. Il s'agit d'un modèle à risque masqué qui a fait cette année l'objet du stage de DEA de Marc Lavarde (université Paris 11 Orsay.) Il a abouti à la réalisation d'une maquette pour laquelle les paramètres de ce modèle peuvent être estimés soit par les algorithmes EM et SEM, soit par approche bayésienne utilisant l'algorithme d'échantillonnage préférentiel BRM. Les résultats obtenus sont très encourageants et la réalisation d'un logiciel en collaboration avec l'IRSN est envisagée.

6.3.3. Réseaux bayésiens et applications à la maintenance

Participants : Gilles Celeux, Franck Corset.

Dans le cadre d'une convention d'étude et de recherche avec le groupe « Sûreté, Diagnostic, Maintenance » de EDF R&D, nous nous sommes intéressés à modéliser les dégradations d'un matériel par un réseau bayésien, prenant en compte de manière simple et explicite les facteurs fonctionnels pouvant influencer l'apparition de maladies sur des systèmes mécaniques. Nous avons proposé une démarche complète pour modéliser le processus de dégradation d'un système mécanique d'une installation nucléaire. Nous nous sommes principalement attachés à la construction des réseaux bayésiens et à leurs analyses dans ce contexte de sûreté de fonctionnement. Le précédent contrat a permis de proposer une procédure conviviale et sûre pour construire le réseau bayésien. Cette procédure ajoute, dans un soucis de simplification, des indépendances conditionnelles via des modèles log-linéaires. Pour l'exploitation des données engendrées par les réseaux bayésiens, nous avons expérimenté les nombreuses méthodes d'analyse basées principalement sur l'analyse de sensibilité. Nous nous sommes attachés à proposer des outils simples de caractérisation des événements saillants contenus dans un réseau. Ainsi, une représentation graphique des scénarios les plus critiques nous a permis d'identifier les variables importantes. D'autre part, nous avons intégré les actions de maintenance comme variables du réseau bayésien. Ainsi, nous avons pu mesurer grâce à une inférence statistique, l'influence des actions de maintenance sur les probabilités de dégradation ou de défaillance du système. De plus, nous avons proposé d'intégrer par une inférence bayésienne utilisant la loi de Dirichlet les données

de retour d'expérience dans le réseau bayésien. Dans ce cadre, nous proposons de quantifier simplement la confiance que l'on attribue dans les avis d'experts par un paramètre de la loi a priori qui peut être traduit en termes de données de retour d'expérience qu'un expert aurait intégré dans sa connaissance. Cette mise en correspondance est très adéquate pour une prise en compte raisonnable et contrôlable des opinions d'expert.

6.3.4. Modélisation d'un changement de comportement de maintenance pour des matériels ayant une période de garantie

Participants : Gilles Celeux, Franck Corset.

Suite à une convention d'étude et de recherche avec le groupe « Sûreté, Diagnostic, Maintenance » de EDF R&D, où nous avons été amené à modéliser un changement de comportement de maintenance pour des systèmes mécaniques de centrales nucléaires, nous nous sommes intéressés à un modèle similaire afin de prendre en compte une période de garantie d'un composant. Cette période de garantie est caractérisée par un comportement de maintenance différent. En effet, lors de la période de garantie, les responsables de maintenance rebutent tous les composants qui présentent la moindre dégradation, ce qui n'est plus le cas lorsque le composant n'est plus garanti. De plus, après la période de garantie, des réparations minimales sont effectuées sans que celles-ci ne soient répertoriées dans les données de retour d'expérience. Ainsi, pour la période de non garantie, un non rebut provient soit d'une vraie censure à droite soit d'une censure à gauche due à une défaillance non signalée. Les données sont doublement censurées. Nous avons proposé, pour ce type de données suivant une loi exponentielle, de modéliser ce facteur humain susceptible d'entacher l'estimation de la loi de durée de vie par un modèle de données incomplètes. Nous avons effectué des simulations qui montrent le bon comportement de l'estimateur du maximum de vraisemblance qui permet de réduire fortement le biais dû à ce comportement humain [36].

6.3.5. Modélisation et estimation de queues de distributions

Participants : Jean Diebolt, Myriam Garrido, Stéphane Girard.

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

Lors d'un premier contrat, il est ressorti que la méthode ET (*exponential tail*) pouvait être un moyen simple de réaliser l'estimation de ces quantiles. Un deuxième contrat a étudié le comportement asymptotique de cette méthode ; ce qui a notamment permis la mise en place d'un test d'adéquation de modèles paramétriques à la queue de distribution. En pratique, il arrive que les tests d'adéquation usuels (dépendant principalement de la partie centrale de la distribution) aboutissent à des conclusions différentes de celles de ce test extrême. Ainsi, un troisième contrat a proposé des procédures de régularisation de la loi obtenue de sorte qu'elle ne perde pas trop de son ajustement central mais qu'elle s'adapte mieux en queue de distribution. La méthode finalement retenue est celle de la régularisation bayésienne qui permet la prise en compte d'un avis d'expert.

Nous avons exploré dans le cadre du contrat [62] une estimation bayésienne des lois de Pareto généralisées, lois limites de la loi des excès quand le seuil tend vers l'infini, qui permettent l'estimation des queues de distribution et des quantiles extrêmes. L'introduction d'une méthode bayésienne (notamment avec un avis d'expert sur les queues de distribution) nous permet de réduire le biais inhérent à la méthode d'estimation des quantiles extrêmes, en particulier lorsque l'on dispose d'un avis d'expert sur la queue de distribution. Certaines propriétés de l'outil bayésien n'ont pas encore été étudiées dans ce cadre, et restent à explorer, par exemple l'utilisation de la loi prédictive a posteriori, ainsi que d'intervalles de crédibilité pour les différents quantités estimées.

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

Participant : Jean-Baptiste Durand.

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

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

Ces hypothèses nous conduisent à considérer que les durées inter-défaillance obéissent à un modèle de Markov caché à lois conditionnelles exponentielles. Des techniques de choix de modèles basées sur des critères de type vraisemblance pénalisée ou validation croisée permettent de déterminer le nombre de versions significatives du logiciel (nombre d'états cachés) ainsi que le type de dynamique dans l'amélioration ou la dégradation du logiciel (matrice de transition). Le modèle est mis en compétition avec des modèles classiques de croissance de fiabilité, sur un critère de capacité prédictive (graphes *uplot*). Ce critère permet de conclure que le modèle est compétitif par rapport aux autres dans le cas de logiciels à faible croissance de fiabilité ou dans le cas où la fiabilité est susceptible de décroître.

6.3.7. Indice de sensibilité

Participants : Julien Jacques, Christian Lavergne.

Débutée en novembre 2001, en collaboration avec le Laboratoire de Conduite et Fiabilité des Réacteurs du CEA de Cadarache la thèse de Julien Jacques a pour objectif d'apporter des éléments de réponse pour évaluer l'impact des incertitudes liées à la simplification et la modification de modèles, dans le cadre d'études de propagation d'incertitudes et d'analyse de sensibilité.

Nous avons plus particulièrement étudié l'analyse de sensibilité par la méthode de Sobol. Elle repose sur une connaissance forte du modèle, des entrées uniformes et indépendantes.

Durant cette année, nous avons étudié le comportement de la méthode de Sobol lorsque l'une de ces hypothèses est relâchée.

Ajout d'une variable par addition ou multipication : nous avons mis en évidence le lien entre l'analyse de sensibilité du nouveau modèle et celle de l'ancien modèle.

Relâchement de l'hypothèse de lois uniformes des entrées : la décomposition de Sobol reste vraie dans le cas d'entrées non uniformes et en particulier gaussiennes. Mais le changement de loi, équivalent à un changement de modèles, influe de façon très significative sur les indices de sensibilité.

Relachement de l'hypothèse d'indépendance : dans le cas d'un modèle à 2 entrées dépendantes, nous avons proposé un nouvel indice qui tient compte de la covariance conditionnelle de la sortie sachant les 2 entrées.

6.4. Statistique biomédicale

Participants : Gilles Celeux, Florence Forbes, Paulo Gonçalvès, Olivier Martin, Christian Lavergne, Christophe Lenoir, Matthieu Vignes, Claudine Robert.

6.4.1. Modèles markoviens parcimonieux pour la biologie moléculaire

Participants : Florence Forbes, Matthieu Vignes.

Travail en collaboration avec Alain Viari, projet Helix, Inria Rhône-Alpes.

Les chaînes de Markov constituent un outil très employé en bioinformatique, notamment pour modéliser les dépendances entre positions sur une molécule d'ADN. Leur usage reste néanmoins limité par un inconvénient majeur : le nombre de paramètres d'un modèle de Markov complet croît exponentiellement avec l'ordre de la chaîne ; en $O(A^k)$ si A est le nombre d'états (i.e. la taille de l'alphabet en biologie) et k l'ordre du modèle. Il semble donc intéressant d'envisager des modèles markoviens plus parcimonieux. Parmi les différentes propositions qui ont été faites, le modèle MTD (Mixture of Transition Distributions) [58], semble fournir, sur des processus markoviens classiques, des résultats très proches des chaînes de Markov complètes. L'économie de paramètres est importante puisque, dans ce cas, le nombre de paramètres croît en $O(A^2 + k)$, c'est-à-dire linéairement avec l'ordre. Le stage de DEA de Matthieu Vignes [57] avait pour objectif d'évaluer la pertinence du modèle MTD (ou de l'un de ses dérivés) à des problématiques biologiques types. Nous avons envisagé plus précisément quatre applications : i) détection de gènes (ceci demande une adaptation du modèle

MTD homogène à un modèle MTD périodique) ; ii) détection d'hélices protéiques transmembranaires (cas des protéines) ; iii) reconnaissance des structures secondaires de protéines (hélices alpha et feuillets beta) et iv) détection des peptides signaux. Les premiers résultats sont plutôt décevants : pour ce qui est de la détection de gènes, il semble qu'il n'y ait pas d'amélioration significative au delà de l'ordre 2.

6.4.2. Analyse de données issues de puces à ADN

Participants : Gille Celeux, Christian Lavergne, Olivier Martin.

Des progrés technologiques récents, comme les puces à ADN [66] par exemple, permettent de mesurer simultanément le niveau d'expression de plusieurs milliers de gènes. Le volume et la complexité des données obtenues rendent les études difficiles et nécessitent des développements statistiques dans plusieurs axes de recherche. Les domaines que nous privilégions concernent la recherche de gènes différentiellement exprimés et la classification de profils d'expresion dans un contexte de forte variabilité. En collaboration avec une équipe de l'IPMC (Institut de Pharmacologie Moléculaire et Cellulaire, Nice) dirigée par Pascal Barbry, nous avons abordé ces deux problématiques en nous efforçant de tirer parti des répétitions pour atténuer les effets de variabilité des mesures. Dans le cadre du stage de biostatistique de Florence Bois-Jover [53], nous avons ainsi étudié les modèles d'analyse de variance pour l'étude des gènes différentiellement exprimés. Nous nous sommes plus particulièrement intéressés au problème du choix de modèle d'analyse de variance.

Mais notre contribution la plus importante a été de proposer de traiter la classification de profils d'expression à partir de données répétées par un modèle de mélange de modèles linéaires mixtes [47]. Ainsi, la classification de données répétées permet d'identifier des groupes de gènes avec des profils d'expression similaires d'une part et de mieux caractériser la dispersion des classes d'autre part [45]. Ce type de modèle est estimé par l'algorithme EM sans difficulté particulière. Il constitue un outil puissant pour la prise en compte d'effets aléatoires, pouvant être de nature très différente, dans un contexte de classification. Son champ d'application dépasse largement celui de la classification de profils d'expression de gènes.

6.4.3. Analyse de données issues du protocole SPARK pour la maladie de Parkinson

Participants : Gilles Celeux, Christian Lavergne, Claudine Robert.

Ce travail se situe dans le cadre d'une collaboration avec les membres du service de neurologie du Pr. Pierre Pollack pour l'analyse statistique des données SPARK. Le protocole SPARK financé par la Fondation pour la Recherche Médicale et dont le promoteur est le CHU de Grenoble a pour objectifs d'étudier l'effet de la stimulation cérébrale profonde du noyau sous thalamique dans la maladie de Parkinson chez 110 patients en quantifiant l'amélioration des scores moteurs, des activités de la vie quotidienne, des états mentaux et comportementaux des patients. Il s'agit d'un projet multicentrique comportant 4 centres Français : Grenoble, Paris Pitié-Salpétrière, Lille et Bordeaux. Le protocole SPARK a été supervisé par Claudine Robert, l'acquisition des données a débuté en novembre 1998 et s'est achevée en mars 2002. L'analyse statistique a fait l'objet du stage de maîtrise de Virginie Roy [56] mettant en œuvre des outils classiques comme l'analyse en composantes principales, l'analyse de la variance et la régression linéaire. Tous les effets significatifs dus à l'effet de la stimulation cérébrale ont été quantifiés et des modèles pour prédire l'amélioration clinique ont été proposés.

Dans un deuxième temps il s'agira de comparer le coût social et économique d'un patient avant et après l'intervention chirurgicale par stimulation cérébrale profonde.

6.4.4. Analyse du rythme cardiaque chez la souris

Participants : Paulo Gonçalvès, Christian Lavergne, Christophe Lenoir.

Ce travail de recherche s'effectue avec Bernard Swynghedauw dans le cadre d'une collaboration avec l'Unité de Recherche U572 de l'INSERM abritée à l'hôpital Lariboisière (Paris), et correspond en partie au travail de thèse de C. Lenoir. Dans cette étude, nous cherchons à mieux identifier le rôle du système nerveux autonome (SNA) dans la variabilité du rythme cardiaque chez la souris. Plus spécifiquement, nous cherchons à mieux caractériser la balance sympathico-vagale liée aux deux branches du SNA, à partir des séries chronologiques de RR (suite des intervalles temporels entre deux battements de coeur consécutifs). Le plan expérimental

utilisé se compose de N individus constituant un échantillon homogène (même souche, même sexe, même age) sur lesquels sont effectuées les mesures suivantes :

- électrocardiogrammes de contrôle (notés intercept)
- électrocardiogrammes enregistrés à 20, 40 et 60 minutes après injection d'un placebo
- électrocardiogrammes enregistrés à 20, 40 et 60 minutes après injection d'atropine (inhibiteur de la branche vagale)
- électrocardiogrammes enregistrés à 20, 40 et 60 minutes après injection de propranolol (β-bloquant de la branche sympathique)
- électrocardiogrammes enregistrés à 20, 40 et 60 minutes après injection simultanée d'atropine et de propranolol (simulation d'un dénervement du myocarde)

Après estimation des séries chronologiques RR correspondantes, les variables que nous cherchons à modéliser sont le rythme cardiaque moyen calculé sur cinq minutes, noté $\bar{RR}$, et la variabilité du rythme cardiaque sur cette même durée et mesurée par la variance $\sigma 2_{RR}$. Les effets que l'on cherche à identifier sont ceux des drogues, et du temps après injection. Pour tenir compte à la fois des mesures répétées et de certaines données manquantes (plan d'expérience non-équilibré), nous utilisons un modèle linéaire généralisé à effet mixte, qui permet d'isoler (et de s'affranchir de) l'*effet souris* sur les mesures, effet qui dans une analyse classique de la variance masque par son énorme intensité toute action liée à l'injection des drogues. À ce stade de notre étude, il semble que le modèle retenu par le critère AIC ne retient que la variable explicative soit le propranolol, dont l'effet β-bloquant s'accentue avec le temps après injection. Contrairement aux autres mammifères étudiés (homme, rat, chien,...), l'atropine est sans effet significatif sur le rythme cardiaque de la souris (outre l'effet *stress* dû à l'injection), ce qui laisse supposer l'absence de tonus vagal chez ce rongeur.

Nous envisageons à court terme l'étude comparative de différents fonds génétiques (souches) de souris, ainsi que l'application de modèles mixtes à d'autres attributs, tels que la régularité, le degré de corrélation à long terme, la puissance spectrale estimés sur les séries de RR.

6.4.5. Modélisation statistique de la plasticité de l'architecture des arbres : analyse de données longitudinales

Participants : Carine Véra, Christian Lavergne.

L'objectif de la thèse de Carine Véra est de développer des méthodes d'analyse de données longitudinales pour l'étude de la croissance (hauteur, longueur des branches, diamètre) d'arbres forestiers en fonction de facteurs environnementaux variant selon le temps (ex : climat), l'espace (ex : station) ou la position dans l'arbre.

Un des premiers objectifs a été de se constituer une base de données. Pour cela, on a construit des séquences multivariées contenant des informations sur la croissance de troncs d'arbres forestiers et ayant un pas de temps annuel. On évalue actuellement une première famille de modèles pour modéliser ces données : la famille des modèles linéaires à effets aléatoires. Une première étape a consisté à modéliser simplement la tendance des données en envisageant leur ajustement par divers modèles linéaires à effets fixes. Le meilleur modèle s'avère être un modèle linéaire avec autant de paramètres que le nombre d'années des arbres étudiés. Ceci souligne la nécessité de prendre en compte d'autres éléments influant directement sur la croissance des arbres. Il faut ainsi mettre en place une approche d'analyse exploratoire des données, afin de détecter l'effet d'évènements climatiques particuliers. Cela permettra de prendre en compte les covariables climatiques dans notre modèle, avec la difficulté supplémentaire liée au changement d'échelle entre la réponse annuelle de l'arbre et le pas de temps journalier des covariables climatiques. C'est un point délicat de nos données qui sont intermédiaires entre données longitudinales (un grand nombre de séquences courtes) et séries chronologiques (une seule séquence longue). L'analyse exploratoire de nos données doit donc combiner des méthodes d'analyse relatives à chacun des deux domaines.

6.5. Inférence statistique pour le traitement du signal et des images

6.5.1. Analyse en composantes principales non-linéaire pour le traitement d'images

Participant : Stéphane Girard.

Travail en collaboration avec Serge Iovleff de l'Université de Bretagne Sud, invité une semaine au sein d'IS2.

Lors de sa thèse, Stéphane Girard a proposé une nouvelle méthode d'analyse en composantes principales (ACP) non linéaire. L'originalité de cette méthode provient de l'approche retenue. Constatant que l'ACP classique construit des approximations de nuages de points par des sous-espaces vectoriels, cette méthode est étendue en construisant des approximations par des variétés [9]. Cette méthode est particulièrement efficace dans le cas où les observations sont des images [61] et sont donc situées dans des espaces de très grande dimension. La définition d'un cadre probabiliste pour cette méthode a fait l'objet d'une collaboration avec Serge Iovleff [50].

6.5.2. Estimation de frontières

Participants : Guillaume Bouchard, Stéphane Girard, Anatoli Iouditski.

Travail en collaboration avec Pierre Jacob, Ludovic Menneteau (Université Montpellier 2) et Alexandre Nazin (IPU, Moscou, Russie).

Cette étude a pour point de départ un article de P. Jacob et C. Suquet, où est exposé un principe d'estimation du support d'un processus de Poisson lorsque son intensité est connue. Le travail mené avec Pierre Jacob [25][24] consiste à étendre ces résultats à l'estimation de support de processus dont on ne connait pas l'intensité en utilisant les statistiques extrêmes. Nous avons également proposé un estimateur basé sur une régression non paramétrique par la méthode du noyau des valeurs extrêmes [67] et établi un principe de grandes déviations [51]. Le travail de Guillaume Bouchard et de Stéphane Girard consiste à reformuler le problème de l'estimation de frontière comme un problème d'optimisation linéaire [30]. Cette recherche est faite en collaboration avec Anatoli Iouditski et Alexandre Nazin qui étudient plus particulièrement les propriétés asymptotiques de l'estimateur ainsi défini.

6.5.3. Analyse de signaux et d'images en modes propres

Participant : Paulo Gonçalvès.

Travail en collaboration avec Patrick Flandrin (CNRS, ENS Lyon) et Paulo Oliveira (IST-ISR, Lisbonne).

L'*Empirical Mode Decomposition*[70] (EMD), est un algorithme itératif qui permet de décomposer un signal complexe (constitué d'un grand nombre de composantes spectrales et/ou fréquentielles) en modes propres de complexité réduite. Dans le cas déterministe, comme dans le cas aléatoire, les propriétés spectrales, respectivement statistiques, des composantes obtenues par cette décomposition ressemblent fortement à celles des décompositions en ondelettes. Toutefois, alors que ces dernières obligent au choix a priori d'une ondelette d'analyse, l'EMD est une approche totalement adaptative (signal dépendante) se comportant comme un filtre à réponse impulsionnelle variable en temps, et donc particulièrement bien adaptée aux non-stationnarités des signaux.

Avec P. Flandrin, nous recherchons une formalisation analytique de l'algorithme, seule définition connue de l'EMD, pour permettre l'analyse théorique de cette décomposition puis l'évaluation de ses performances. En marge de ce travail de conceptualisation, nous travaillons également au développement de versions plus robustes de l'algorithme EMD, ainsi qu'au développement d'une version en ligne.

Enfin, avec P. Oliveira, nous travaillons sur une adaptation de l'EMD aux images. Les problèmes posés sont de plusieurs natures : définition d'extrema locaux en 2D, choix d'une interpolation de surfaces, critères d'arrêt (locaux versus globaux)...Une application visée de l'EMD-2D est le repérage de structures (types vortex) localisées dans des champs d'écoulements turbulents.

6.5.4. Test d'existence de moments

Participant : Paulo Gonçalvès.

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

Nous avons poursuivi l'étude sur le test d'existence de moments d'une variable aléatoire proposé dans [68]. Notamment, nous avons établi un résultat théorique qui met en relation la vitesse de décroissance des queues de distributions, la régularité locale de la fonction caractéristique et les bornes $\lambda^-, \lambda^+)$ de l'intervalle sur lequel tous les moments d'ordre r existent. Ce théorème constitue un socle théorique fort qui :

1. généralise à l'ensemble des réels, des résultats qui n'étaient valables jusqu'alors que pour les moments d'ordre inférieurs à deux ;
2. garantit la validité de l'estimateur empirique proposé.

Ces résultats nouveaux font l'objet d'un rapport de recherche Inria [52].

6.5.5. Estimation de lois stables

Participant : Paulo Gonçalvès.

Travail en collaboration avec Anestis Antoniadis (UJF-IMAG) et Andrey Feuerverger (Université de Toronto, Canada).

Les distributions α-stables sont caractérisées par un jeu θ de quatre paramètres que l'on peut estimer par des procédures approchées du maximum de vraisemblance, et souvent très coûteuses en temps de calcul (Koutrouvelis, McCulloch, Nollan,...). Nous nous intéressons ici à une méthode originale d'estimation de θ, par régression non linéaire de la décomposition en ondelettes de la fonction caractéristique $\Phi_\theta(t)$ d'une variable stable. La projection sur une base d'ondelettes, fonctions oscillantes à décroissance rapide (voire même à support compact), produit une matrice creuse de coefficients $d_{j,k}$, dont les valeurs non nulles se concentrent essentiellement au voisinage des points de singularité de la fonction analysée. L'unique point de discontinuité de $\Phi_\theta(t)$ étant situé en $t = 0$, les seuls coefficients d'ondelette $d_{j,k}$ significatifs se localisent autour de l'origine. Cela permet finalement de ne retenir que très peu de points (au minimum un par échelle) dans le schéma de régression non linéaire utilisé pour estimer θ.

6.5.6. Diffusion de représentations temps-fréquence pour un problème décisionnel

Participant : Paulo Gonçalvès.

Travail en collaboration avec Julien Gosme (Université de Technologie de Troyes) dans le cadre de sa thèse.

À l'issue de son stage de DEA, Julien Gosme a poursuivi dans le cadre d'une thèse de doctorat, son travail sur la diffusion adaptative des représentations temps-fréquence [69] appliquées au problème de détection non paramétrique. L'idée séminale de ces représentations est de diminuer par lissage adapté la complexité du détecteur mesurée par la dimension de Vapnik-Chervonenkis. Des résultats encourageants ont été obtenus dans cette direction et ont fait l'objet d'un article de conférence [44].

7. Contrats industriels

7.1. Etude de courbes de consommation électrique

Participants : Gilles Celeux, Jean-Baptiste Durand, Myriam Garrido.

Ce contrat de type CRECO avec le département « clientèle » de EDF-DRD CLAMART a pour objet l'analyse statistique de courbes de consommation électrique. L'hypothèse que la consommation d'un ménage dépend d'un état non observé lié au type d'activité (repas, veillée, sommeil, etc.) et que ces états obéissent à un régime markovien nous conduisent à modéliser les courbes de consommation par des chaînes de Markov cachées.

La première étape de l'analyse consiste à estimer et à supprimer tous les effets non aléatoires tels que la saisonnalité et les effets dus à l'heure de la journée ou au type de contrat, par une analyse de la variance. Les résidus sont alors modélisés par une chaîne de Markov cachée. Les états cachés sont restaurés pour leur interprétation vis-à-vis des différents usages d'appareils électriques. Pour ce faire, on construit un tableau de contingence mettant en relation les usages réels, disponibles pour les données d'apprentissage, et les états cachés restaurés. L'analyse factorielle des correspondances du tableau de contingence permet la visualisation de différents niveaux de consommation électrique, associés à des groupes d'usages.

L'un des intérêts du modèle est de permettre, en utilisant le tableau de contingence, l'estimation des usages lorsque ceux-ci sont inconnus, à partir de la consommation totale. Cette estimation se base sur la restauration des états cachés puis sur l'affectation de la consommation aux différents usages, proportionnellement à leur part respective donnée par le tableau de contingence. Ce mode de restauration aboutissant à des résultats

parfois peu réalistes (proportionnalité de la consommation des différents appareils électriques), nous avons pris en compte la consommation type due à chaque usage pour pondérer ces estimateurs. Ces différentes procédures de restauration d'usage ont été intégrées à un logiciel destiné à EDF, qui permet la visualisation des usages restaurés.

La mise en œuvre des modèles de Markov cachés requiert le choix d'un nombre d'états cachés. Vu le nombre de données disponibles et le manque de réalisme de l'hypothèse de Markov cachée, les critères de sélection de modèles utilisés (BIC, ICL et la validation croisée) conduiraient à choisir des modèles dont la grande complexité rendrait impossible l'interprétation vis-à-vis des usages. C'est pourquoi nous avons opté pour un critère prenant en compte cet objectif d'interprétation : il s'agit de maximiser l'écart à l'indépendance entre les usages et les états cachés. Ce critère n'est cependant pas très satisfaisant dans la mesure où il ne prend pas en compte l'adéquation entre les données et le modèle choisi ; c'est pourquoi nous étudions un critère prenant en compte à la fois l'adéquation (mesurée par la vraisemblance) et la classifiabilité des observations vis-à-vis d'une partition a priori des usages (mesurée par un critère de type entropique présentant des analogies avec le critère ICL [2]).

7.2. Utilisation des réseaux bayésiens en fiabilité

Participants : Gilles Celeux, Franck Corset.

Ce contrat de type CERD avec le département « Surveillance, Diagnostic, Maintenance » de EDF R&D concernait l'utilisation de réseaux bayésiens en fiabilité et en sûreté de fonctionnement. Cette étude faisait suite à une étude de même type effectuée les deux dernières années. Cette année nous nous sommes concentrés sur l'analyse statistique des données issues des réseaux bayésiens. Nous proposons des indices simples basés sur un rapport de vraisemblance. Ce score nous permet de classer les scénarios possibles, suivant leurs influences sur la variable d'intérêt, ici la dégradation et la défaillance du joint 1 d'une pompe primaire 900 MW. Une fois ces variables importantes identifiées, l'aide à la décision porte sur l'intégration des actions de maintenance sur ces variables jugées importantes. Pour cela, nous considérons les tâches de maintenance comme de nouveaux nœuds du réseau bayésien. Cela permet notamment de prendre en compte le gain apporté par cette action de maintenance grâce à de nouvelles probabilités conditionnelles. Ces actions de maintenance ainsi que leurs gains et leurs coûts sont donnés par les experts. L'impact de ces actions de maintenance est mesuré par une inférence statistique sur ce nouveau réseau prenant en compte les actions de maintenance. Enfin, un logiciel a été réalisé grâce à la boite à outil BNT de Matlab.

7.3. Scénarios de défaillance de pénétrations de fond de cuves (PFC)

Participants : Guillaume Bouchard, Gilles Celeux.

Ce contrat de type CRECO avec le groupe Fiabilité des composants et structures de EDF R&D CHATOU commencé en 2001 avait pour objectif la modélisation bayesienne des fuites dans les fonds de cuves de réacteurs nucléaires. Ce problème est particulier car aucune défaillance (c'est-à-dire ici une fissuration) n'a été observée. Les données disponibles sont des scénarios hypothétiques de défaillance ainsi que des données issues de connaissances a priori sur le modèle (par exemple la température à l'intérieur des cuves).

En sortie, nous obtenons une loi a posteriori pour les paramètres du modèle décrivant les fissurations. Cette réactualisation des paramètres est ensuite utilisée pour estimer le nombre de défaillances sur les cuves non contrôlées.

Le travail effectué s'est effectué en plusieurs étapes :

- définition d'un modèle d'amorçage des fissures,
- estimation des paramètres a posteriori par une méthode de de chaîne de Markov de Monte-Carlo,
- prise en compte du temps de propagation des fissures,
- exploitation des résultats.

Ce travail [23] a donné lieu à la création d'un programme en C++ pour estimer les lois a posteriori, ainsi que des fonctions Matlab qui exploitent les résultats.

7.4. Contrat edf sur les queues de distribution de probabilité

Participants : Jean Diebolt, Jérôme Ecarnot, Myriam Garrido, Stéphane Girard.

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

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

L'étude menée dans le premier contrat prolonge le travail des trois années précédentes sur ce même thème. Nous disposons maintenant de tests permettant de vérifier l'adéquation d'un modèle paramétrique à un échantillon, tant du point de vue de sa forme globale que du point de vue de sa queue de distribution, ainsi que d'une procédure de régularisation bayésienne qui nous permet de construire un modèle global. Nous explorons la piste bayésienne pour l'estimation des lois de Pareto g/'en/'eralisées (voir paragraphe 6.3.5). Une Journée EDF de formation à la maquette logiciel Extremes a été organisée en mai 2002. Le second contrat concerne la programmation d'un logiciel pour la modélisation des évènements rares. Il s'agit de reprendre et d'étendre la maquette logiciel Extremes.

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

Participants : Christian Lavergne, Gérard Boudjema.

Une étude de sensibilité du code de calcul d'estimation des flux neutroniques reçus par les cuves de réacteurs (code STAY'SL) à été menée dans le cadre du contrat de collaboration avec le CEA de Cadarache (DER-STR/LCFR). Après une phase préparatoire d'analyse du processus de préparations des données (menée en 2001), la partie calculs a été expertisée, des corrections on été proposées et une analyse de l'effet sur les sorties des incertitudes sur les entrées été effectuée. Les indices de Sobol ont été ensuite calculés de manière à mettre en évidence les variables qui influencent le plus les sorties du modèle. Les codes Matlab de calcul et de simulations ont été fournis au CEA.

8. Actions régionales, nationales et internationales

8.1. Actions régionales

IS2 participe régulièrement au séminaire de statistique du LMC-SMS à Grenoble et G. Celeux est l'un des organisateurs. Dans ce cadre, plusieurs conférenciers ont été invités.

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

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

P. Gonçalvès est membre de l'action IMAG « Analyse Multirésolution, Ondelettes et Applications », dirigée par Valérie Perrier, Professeur à l'Ensimag (INPG).

Le groupe FIMA qui a fait l'objet d'un soutien par une ARC locale a été créé. Ce groupe a pour but de fédérer ces activités de recherche entre le LMC et l'INRIA pour d'une part renforcer la visibilité du pôle grenoblois de recherche en fiabilité, et d'autre part développer de nouveaux axes de recherche. En particulier, nous souhaitons développer les relations entre les deux organismes participants et tous les partenaires locaux intéressés par la sûreté de fonctionnement, aussi bien les laboratoires de recherche que les entreprises. La

principale activité de FIMA est le groupe de travail qui se réunit approximativement une fois par mois et auquel les membres d'IS2 impliqués en fiabilité participent activement : cette année Gilles Celeux, Franck Corset (deux fois), Jean-Baptiste Durand et Myriam Garrido ont exposé à ce groupe de travail.

8.2. Actions nationales

P. Gonçalvès entretient une collaboration régulière avec l'équipe U572 de l'Inserm à l'hôpital Lariboisière (Paris), et dans ce cadre il co-encadre la thèse de Christophe Lenoir sur l'analyse du rythme cardiaque chez la souris.

P. Gonçalvès fait partie du groupe de travail « Analyse et modélisation de la régularité des images naturelles », financé par le GdR-PRC ISIS (CNRS).

La collaboration avec Yann Guédon du Cirad concerne des recherches sur l'analyse de données longitudinales dans le cadre de la structure exponentielle. Ce thème fait l'objet de la thèse de Carine Véra (ASC INRA), co-encadrée par Yann Guédon et Christian Lavergne. D'autre part cette collaboration s'est concrétisée par le détachement de Yann Guédon à l'Inria Rhône-Alpes dans le projet IS2. Sa recherche durant son détachement portera sur les modèles markoviens (cachés ou non), les modèles graphiques et la sélection de modèles.

8.3. Relations bilatérales internationales

8.3.1. Europe

Mike Titterington de l'université de Glasgow a passé un mois lors de deux séjours en avril et en octobre dans le projet IS2. Il a entamé des recheches avec Gilles Celeux et Florence Forbes sur la sélection de modèles dans un contexte bayésien (voir 6.2.3, 6.2.2).

G. Celeux poursuit sa collaboration avec le LEAD de l'université de Lisbonne, et il a donné un séminaire sur les mélanges de modèles linéaires mixtes dans cette université.

P. Gonçalvès et G. Celeux poursuivent leur coopération scientifique bilatérale INRIA/ICCTI (Portugal). L'Institut des Systèmes et Robotique de l'Institut Supérieur de Technologie (Lisbonne) est notre collaborateur au Portugal, et avec lui nous menons une étude sur le thème *Inférences statistiques en Traitement du Signal : Mesures spectrales instantanées et modèles de mélanges gaussiens*. Les recherches menées portent sur un modèle de chaîne de Markov cachée pour un problème de suivi de dynamique (voir 6.1.5). D'autre part, Gilles Celeux est rapporteur de la thèse de Jacinto Nascimento soutenue en janvier 2003. Nous arrivons maintenant au terme des deux premières années de financement, et avons répondu à l'appel d'offre 2003, pour une demande de renouvellement.

8.3.2. Maghreb

G. Celeux poursuit des relations de recherche régulières avec A. Mkhadri (université de Marrakech) sur la sélection de modèles par validation croisée.

8.3.3. Amérique du Nord

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

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

8.4. Accueil de chercheurs étrangers

Mike Titterington, université de Glasgow, a effectué deux séjours de quinze jours dans le projet IS2.

Alexandre Nazin (IPU, Moscou, Russie) a effectué un séjour d'un mois dans le projet IS2.

Mohamed Saidane (université de Tunis) a effectué un séjour d'un mois.

9. Diffusion des résultats

9.1. Animation de la communauté scientifique

P. Gonçalvès est Éditeur Associé de *IEEE Signal Processing Letters*.

P. Gonçalvès est correspondant IS2 pour l'Action Spécifique STIC du CNRS : "Lois d'échelle, modèles et outils".

9.2. Enseignement universitaire

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

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

S. Girard est, depuis le 1er février 2002, Maître de conférences à l'université Joseph Fourier.

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

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

9.3. Participation à des colloques, séminaires, invitations

G. Celeux a été conférencier invité au *workshop Mixture Models between Theory and Applications* de l'université La Sapienza de Rome en septembre et au colloque Franco-Libanais de statistique qui s'est tenu à Beyrouth en septembre.

G. Celeux, J.-B. Durand, M. Garrido, S. Girard, C. Lavergne, et O. Martin ont participé aux XXXIIIèmes journées de statistique de la SfdS, à Bruxelles en mai 2002.

G. Celeux, F. Corset, J. Diebolt et M. Garrido ont participé au congrès $\lambda\mu$ 13, ESREL 2002, Aide à la décision et maîtrise des risques, à Lyon en mars 2002.

G.Celeux, J. Diebolt et M. Garrido ont participé aux quatrièmes journées MAS, Modélisation Aléatoire et Statistique, à Grenoble en septembre 2002.

G.Celeux et G. Bouchard ont participé à la conférence Statistical Learning, Theory and Applications, au CNAM Paris en novembre 2002.

F. Forbes a été invitée pour un séminaire à l'université technologique de Munich et à l'Inria Sophia-Antipolis.

F. Forbes, O. Martin et M. Vignes ont participé au seminaire Algo-Bio à Lyon en octobre 2002.

P. Gonçalvès a été conférencier invité au *2nd meeting of the European Study Group of Cardiovascular Oscillations*, Sienne (Italie), avril 2002.

10. Bibliographie

Bibliographie de référence

[1] M. BACHA, G. CELEUX, E. IDÉE, A. LANNOY, D. VASSEUR. *Estimation de modèles de durées de vie fortement censurés.* Eyrolles, Paris, 1998.

[2] C. BIERNACKI, G. CELEUX, G. GOVAERT. *Assessing a mixture model for clustering with the integrated completed likelihood.* in « IEEE Trans. on PAMI », 2000, pages 267-272.

[3] G. CELEUX, J. DIEBOLT. *A stochastic approximation type EM algorithm for the mixture problem.* in « Stochastic and Stochastics Reports », volume 41, 1992, pages 119-134.

[4] G. CELEUX, F. FORBES, N. PEYRARD. *EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation.* in « Pattern Recognition », numéro 1, volume 36, 2003, pages 131-144.

[5] G. CELEUX, G. GOVAERT. *Gaussian parsimonious clustering models.* in « Pattern Recognition », volume 28, 1995, pages 781-793.

[6] P. FLANDRIN, P. GONÇALVÈS, P. ABRY. *Lois d'échelle, Fractales et Ondelettes.* édition Hermès Sciences Publications, série Traité Information - Commande - Communication, volume 2, Abry, P. and Gonçalvès, P. and Lévy Véhel, J. eds, Paris, France, 2002, chapitre Analyses en ondelettes et lois d'échelle.

[7] F. FORBES, A. E. RAFTERY. *Bayesian Morphology : Fast Unsupervised Bayesian Image analysis.* in « Journal of the American Statistical Association », numéro 446, volume 94, June 1999, pages 555-568.

[8] O. FRANÇOIS, C. LAVERGNE. *Design for Evolutionary Algorithms - A Statistical Perspective.* in « IEEE Transactions on Evolutionary Computation », volume 5, 2001, pages 129-148.

[9] S. GIRARD. *A nonlinear PCA based on manifold approximation.* in « Computational Statistics », volume 15(2), 2000, pages 145-167.

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

Livres et monographies

[11] éditeurs P. ABRY, P. GONÇALVÈS, J. E. LÉVY VÉHEL., *Lois d'échelle, Fractales et Ondelettes, volumes 1 et 2.* série Traité Information - Commande - Communication, Hermès Sciences Publications, Paris, France, 2002.

Thèses et habilitations à diriger des recherche

[12] F. CORSET. *Aide à l'optimisation de maintenance à partir de réseaux bayésiens et fiabilité dans un contexte doublement censuré.* thèse de doctorat, Université Joseph Fourier, Grenoble, 2002, Soutenance en janvier 2003.

[13] J.-B. DURAND. *Modèles à structure cachée : inférence, sélection de modèles et applications.* thèse de doctorat, Université Joseph Fourier, Grenoble, 2002, Soutenance en janvier 2003.

[14] M. GARRIDO. *Prédiction des événements rares et estimation des quantiles extrêmes, méthodes de sélection de modèles pour les queues de distribution.* thèse de doctorat, Université Joseph Fourier, 2002.

[15] O. MARTIN. *Approches statistiques pour l'analyse de données de puces à ADN.* thèse de doctorat, Université Joseph Fourier, Grenoble, 2002.

Articles et chapitres de livre

[16] C. BIERNACKI, G. CELEUX, G. GOVAERT. *Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models.* in « Computational Statistics and Data Analysis », 2002, à paraître.

[17] G. CELEUX. *Analyse discriminante.* éditeurs G. GOVAERT., in « L'analyse des données », série Information-Commande-Communication, Hermes Science, 2002, à paraître.

[18] G. CELEUX, F. CORSET, M.-A. GARNERO, C. BREUILS. *Accounting for inspection errors and change in maintenance behaviour.* in « Journal of Management Mathematics Special Issue », numéro 1, volume 13, 2002, pages 51-59.

[19] G. CELEUX, F. FORBES, N. PEYRARD. *EM Procedures Using Mean Field-Like Approximations for Markov Model-Based Image Segmentation.* in « Pattern Recognition », numéro 1, volume 36, 2002, pages 131-144.

[20] J.-B. DURAND, P. GONÇALVÈS, Y. GUÉDON. *Statistical Inference for Hidden Markov Tree Models and Application to Wavelet Trees.* in « IEEE Transactions on Signal Processing », 2002, à paraître.

[21] P. FLANDRIN, P. GONÇALVÈS, P. ABRY. éditeurs P. ABRY, P. GONÇALVÈS, J. LÉVY VÉHEL., *Lois d'échelle, Fractales et Ondelettes.* édition Hermès Sciences Publications, série Traité Information - Commande - Communication, volume 2, Paris, France, 2002, chapitre Analyses en ondelettes et lois d'échelle.

[22] F. FORBES, N. PEYRARD. *Hidden Markov Random Field Model Selection Criteria based on Mean Field-like Approximations.* in « in IEEE trans. PAMI », 2002, à paraître.

[23] G. C. G. BOUCHARD. *Réactualisation des paramètres du temps de fissuration des PFC.*. Novembre, 2002.

[24] S. GIRARD, P. JACOB. *Extreme values and Haar series estimates of point process boundaries.* in « Scandinavian Journal of Statistics », 2002, à paraître.

[25] S. GIRARD, P. JACOB. *Projection estimates of point processes boundaries.* in « Journal of Statistical Planning and Inference », 2002, à paraître.

[26] A. JUDITSKY, O. LAMBERT-LACROIX. *On nonparametric confidence set estimation.* in « Math. Meth. of Stat », 2002, à paraître.

[27] A. JUDITSKY, A. NEMIROVSKI. *On Nonparametric Tests of Positivity / Monotonicity / Convexity.* in « Ann. of Stats », 2002, à paraître.

Communications à des congrès, colloques, etc.

[28] C. BAUBY, D. LAGRANGE, J. DIEBOLT, M. GARRIDO. *Goodness-of-fit test for distribution tails and Bayesian regularization procedure to estimate low probabilities for system decision.* in « $\lambda\mu$ 13, ESREL 2002, Aide à la décision et maîtrise des risques », Lyon, mars, 2002.

[29] G. BOUCHARD. *Mixture of Regressions with Normal Regressors.* in « GfKl 2002, Manheim », 2002.

[30] G. BOUCHARD, S. GIRARD. *Support Vector Learning for Frontier Estimation.* in « Colloque L'Apprentissage Statistique, Théorie et Applications », pages 66-69, CNAM, Paris, novembre, 2002.

[31] G. CELEUX. *Mixture of linear mixed models, application to repeated data clustering.* in « Worshop Mixture Models between Theory and Applications », Roma, septembre, 2002.

[32] G. CELEUX. *Panorama sur les modèles de mélanges.* in « Premier colloque Franco-Libanais de Statistique », Beyrouth, septembre, 2002.

[33] G. CELEUX, J.-B. DURAND. *Choosing the order of a hidden Markov chain through cross-validated likelihood.* in « Compstat2002. Berlin (Allemagne), 24-28 Août », 2002.

[34] F. CORSET, G. CELEUX, A. LANNOY, B. RICARD. *Bayesian Networks as a decision tool in maintenance with expert judgement.* in « ESReDA 23rd Seminar on decision analysis », Delft University, Pays-Bas, 18-19 novembre, 2002.

[35] F. CORSET, G. CELEUX, A. LANNOY, B. RICARD. *Designing a Bayesian Network for preventive maintenance from expert opinions in a rapid and reliable way.* in « ESREL 2002-$\lambda\mu$ 13 », Palais des congrès, Lyon, 18-21 mars, 2002.

[36] F. CORSET. *Modelling a guarantee period and a change of maintenance behaviour.* in « ESReDA 22nd Seminar on maintenance management and optimisation », Madrid, Espagne, 27-28 mai, 2002.

[37] N. DEVICTOR, M. MARQUES, S. BOULÈGUE, P. LAMAGNÈRE, M.-P. VALETA, M. EID, M. GARRIDO. *Analyse statistique de la ténacité : utilisation de méthodes de Monte-Carlo pour estimer l'incertitude sur la distribution de la ténacité.* in « $\lambda\mu$ 13, ESREL 2002, Aide à la décision et maîtrise des risques », Lyon, mars, 2002.

[38] J.-B. DURAND, O. GAUDOIN. *Modélisation du processus des défaillances et corrections d'un logiciel par des chaînes de Markov cachées.* in « XXXIV-èmes Journées de Statistique », pages 212, Bruxelles et Louvain-la-Neuve (Belgique), mai, 2002.

[39] F. FORBES. *Approximation tools for statistical inference in models with intractable structure.* in « Workshop on model-based clustering », University of Washington, Seattle, Etats-Unis, 22-26 juillet, 2002.

[40] M. GARRIDO, J. DIEBOLT, S. GIRARD. *Une nouvelle approche bayésienne pour l'estimation des paramètres d'une loi GPD.* in « XXXIV-èmes Journées de Statistique », Bruxelles et Louvain-la-Neuve (Belgique), mai, 2002.

[41] M. GARRIDO, J. DIEBOLT, S. GIRARD. *Une nouvelle approche bayésienne pour estimer les paramètres d'une loi GPD.* in « quatrièmes journées MAS, Modélisation Aléatoire et Statistique », Grenoble, septembre, 2002.

[42] S. GIRARD, S. IOVLEFF. *Modèles Auto-Associatifs et Analyse en Composantes Principales généralisée.* in « XXXIV-èmes Journées de Statistique », pages 231, Bruxelles et Louvain-la-Neuve (Belgique), mai, 2002.

[43] S. GIRARD, L. MENNETEAU. *Théorèmes limites pour l'estimation du contour d'un processus de Poisson.* in « XXXIV-èmes Journées de Statistique », pages 232, Bruxelles et Louvain-la-Neuve (Belgique), mai, 2002.

[44] J. GOSME, P. GONÇALVÈS, C. RICHARD, R. LENGELLÉ. *Adaptive Diffusion and Discriminant analysis for Complexity Control of Time-Frequency Detectors.* in « XI European Signal Processing Conference », Toulouse (France), Septembre, 2002.

[45] O. MARTIN, G. CELEUX, C. LAVERGNE. *Classification de données répétées issues de puces à ADN. Application à l'analyse de profils d'expression.* in « JOBIM, 2002 », pages 81-92, 2002.

[46] O. MARTIN, G. CELEUX, C. LAVERGNE. *Mélange de modèles linéaires mixtes pour la classification de données répétées. Application à l'analyse de données issues de biopuces.* in « XXXIV-èmes Journées de Statistique », Bruxelles et Louvain-la-Neuve (Belgique), mai, 2002.

Rapports de recherche et publications internes

[47] G. CELEUX, C. LAVERGNE, O. MARTIN. *Mixture of linear mixed models. Application to repeated data clustering.* rapport technique, numéro RR-4566, INRIA, 2002, http://www.inria.fr/rrrt/rr-4566.html.

[48] J. DIEBOLT, M. GARRIDO, G. S.. *Asymptotic normality of the ET method for extreme quantile estimation. Application to the ET test.* rapport technique, numéro 4551, Inria Rhône-Alpe, septembre, 2002, http://www.inria.fr/rrrt/rr-4551.html.

[49] J.-B. DURAND, P. GONÇALVÈS, Y. GUÉDON. *Statistical Inference for Hidden Markov Tree Models and Application to Wavelet Trees.* rapport technique, numéro RR-4248, INRIA, 2002, http://www.inria.fr/rrrt/rr-4248.html, En révision pour *IEEE Transactions on Signal Processing*.

[50] S. GIRARD, S. IOVLEFF. *Auto-Associative Models and Generalized Principal Component Analysis.* rapport technique, numéro 4364, 2002, http://www.inria.fr/rrrt/rr-4364.html.

[51] S. GIRARD, L. MENNETEAU. *Limit theorems for extreme value estimates of point processes boundaries.* rapport technique, numéro 4366, Inria Rhône-Alpes, 2002, http://www.inria.fr/rrrt/rr-4366.html.

[52] P. GONÇALVÈS, R. RIEDI. *Diverging Moments and Parameter Estimation.* rapport technique, numéro RR-4647, INRIA, Novembre, 2002, http://www.inria.fr/rrrt/rr-4647.html.

Divers

[53] F. BOIS-JOVER. *Analyse statistique de données de biopuces.* 2002.

[54] C. LANGLOIS. *Un modèle de durée de vie.* 2002.

[55] C. MARTIN. *Modèles à structure de covariance pour la classification de données spatiales.* Mémoire de 2eme année ENSIMAG, 2002.

[56] V. ROY. *Analyse statistique d'une base de données concernant des patients atteints de la maladie de Parkinson.* Mémoire de maîtrise, 2002.

[57] M. VIGNES. *Modèles de Markov parcimonieux pour la biologie moléculaire.* Mémoire de DEA de mathématiques. UCB Lyon I, 2002.

Bibliographie générale

[58] A. BERCHTOLD, A. RAFTERY. *The Mixture Transition Distribution (MTD) Model for High-Order Markov Chains and Non-Gaussian time Series.* rapport technique, numéro 360, Department of Statistics, University of Washington, Seattle, August 1999.

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

[60] J. BESAG. *On the statistical analysis of dirty pictures.* in « Journal of the Royal Statistical Society, series B », volume 48, 1986, pages 259-302.

[61] B. CHALMOND, S. GIRARD. *Nonlinear modeling of scattered multivariate data and its application to shape change.* in « I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence », numéro 5, volume 21, 1999, pages 422-432.

[62] J. DIEBOLT, C. BAUBY, M. GARRIDO. *Estimation bayésienne de la loi GPD, loi asymptotique des excès au-delà d'un seuil.* 2001, rapport final de convention de recherche Inria-EDF.

[63] F. FORBES, C. FRALEY, D. GEORGIAN-SMITH, D. GOLDHABER, N. PEYRARD, A. RAFTERY. *Region-Of-Interest Selection and Statistical Analysis of Dynamic Breast Magnetic Resonance Imaging Data.* rapport technique, numéro 4249, Inria Rhône-Alpes, 2001, http://www.inria.fr/rrrt/rr-4249.html.

[64] E. GASSIAT. *Likelihood ratio inequalities with application to various mixtures.* 2002, Annales de l'institut Poincaré (à paraître).

[65] S. GEMAN, D. GEMAN. *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images.* in « I.E.E.E. Transactions on Pattern Analysis and Machine Intelligence », volume 6, 1984, pages 721-741.

[66] *The Chipping Forecast.* volume 21, 1999.

[67] S. GIRARD, P. JACOB. *Extreme values and kernel estimates of point processes boundaries.* rapport technique, numéro 0102, ENSAM-INRA-Université Montpellier II, 2001, http://www.math.univ-montp2.fr/probastat/RR.html.

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

[69] P. GONÇALVÈS, E. PAYOT. *Diffusion equation for time frequency representation.* in « Proc. IEEE Digital signal processing workshop », 1998.

[70] N. E. HUANG, ET AL.. *The Empirical Mode Decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis.* in « The Royal Society », 1998.

[71] G. MCLACHLAN, T. KRISHNAM. *The EM algorithm and extensions.* John Wiley, New York, 1997.

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

[73] A. PETTITT, I. WEIR, A. G. HART. *A conditional Autoregressive Gaussian Process for irregularly Spaced Multivariate Data with Application to Modelling Large Sets of Binary Data.* rapport technique, Queensland university of Technology, school of Mathematical sciences, Brisbane, Australia, 2001.

[74] K. ROEDER, L. WASSERMAN. *Practical Bayesian density estimation using mixtures of normals.* in « Journal of the American Statistical Association », volume 92, 1997, pages 894-902.

[75] D. J. SPIEGELHALTER, ET AL.. *Bayesian measures of model complexity and fit.* in « Journal of the Royal Statistical Society, series B », volume 64, 2002, pages 1-34.