Avant-projet IS2

Inférence statistique pour l'industrie et la santé

Localisation : *Grenoble*

Mots-clés : Modélisation statistique, données incomplètes, hétéroscédasticité, modèles linéaires généralisés, principe de maximum d'entropie, algorithmes stochastiques, aide au diagnostic, analyse de durées de vie, événements rares.

1 Composition de l'équipe

Responsable scientifique

Gilles Celeux, DR Inria

Personnel des établissements partenaires

Christian Lavergne, professeur, université Paul Valéry, Montpellier
Claudine Robert, professeur, université Joseph Fourier Grenoble 1

Chercheurs post-doctorants

Mostafa Bacha, boursier Inria
Mhamed-Ali El Aroui, boursier Inria (depuis le 1/10/96)
Joseph Ngatchou Wandji, boursier Inria (depuis le 1/4/96)

Chercheurs doctorants

Henri Bertholon, enseignant CNAM, (depuis le 1/9/96)
Catherine Trottier, allocataire MRT, INPG
Véronique Venditti, boursière Inria
Yann Vernaz, vacataire université Pierre Mendès-France (depuis le 1/10/96)

Collaborateurs extérieurs

Jean-Luc Bosson, praticien hospitalier, CHU de Grenoble
Jean Diebolt, DR CNRS LMC-SMS
Edwige Idée, MC, université de Savoie

2 Présentation du projet

L'avant-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 intrinséquement une partie de l'information nécessaire à l'identification du phénomène étudié est manquante. Ces modèles sont courants (durées de vie censurées, modèles hétéroscédastiques, ...) et puissants (modèles à structure cachée, ...). Ils apparaissent dans de nombreux problèmes statistiques qui se posent en milieu bio-médical 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 (montage industriel compliqué, existence d'une structure de dépendance temporelle ou spatiale, nombreuses variables en jeu, ...). De ce fait, ces modèles sont en général faiblement identifiables en ce sens que, au vu des observations effectivement recueillies, plusieurs jeux différents de paramètres peuvent apparaître également bons. Cela se traduit par une multiplicité des *extrema* locaux des fonctions de contrastes 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 buts de IS2 est de proposer des méthodes d'estimation et d'évaluation efficaces de ces modèles. En particulier, nous nous intéressons à l'utilisation et à l'étude théorique d'algorithmes stochastiques (versions stochastiques de l'algorithme EM, algorithmes MCMC, algorithmes de rééchantillonnage) dont le dénominateur commun est la restauration des données manquantes par simulation.

Les modèles considéreés par IS2 sont par ailleurs souvent dictés par les problèmes qui nous sont soumis. Ainsi le choix de modèles bayésiens pour des problèmes d'analyse de défaillance s'explique 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 récent pour la modélisation des événements rares vient d'un problème que nous a soumis EDF. Dans ce souci de proposer des modèles bien reliés aux préoccupations des utilisateurs, nous nous intéressons sérieusement à une stratégie de modélisation par un principe de maximum d'entropie. Ce principe produit à la fois une relecture éclairante de modèles classiques, et fournit aussi les moyens d'obtenir, plus ou moins automatiquement, des modèles réalistes à partir de quantités jugées pertinentes par les utilisateurs.

La validation des modèles construits et identifiés est bien sûr un élément important de la recherche. Nous l'abordons par des tests non paramétriques ou, dans une perspective bayésienne, par le calcul de critères de parcimonie.

3 Actions de recherche

3.1 Le principe de maximum d'entropie en modélisation statistique

Participants : Gilles Celeux, Claudine Robert, Véronique Venditti

Le principe de maximum d'entropie (PME) renverse la présentation classique de la modélisation statistique au sens où il choisit en premier lieu les quantités statistiques que l'on juge essentielles pour résumer l'information apportée par un jeu de données. Le modèle, c'est-à-dire la loi de probabilité décrivant le phénomène aléatoire, n'apparaît qu'après et doit vérifier des contraintes mettant en jeu ces quantités statistiques essentielles.

La loi du maximum d'entropie est obtenue par maximisation d'une certaine fonctionnelle sur l'ensemble des lois pouvant servir de modèle. La fonctionnelle que nous avons choisie est l'entropie de Shannon, car elle seule permet d'atteindre une loi qui possède la propriété de concentrer les lois empiriques dans son voisinage.

Nous avons montré, que s'il existe un état de Gibbs répondant aux contraintes, celui-ci représente l'unique loi qui réalise le maximum de cette entropie. Cette propriété permet d'éviter le recours habi-

tuel à la méthode des multiplicateurs de Lagrange et autorise ainsi de traiter le cas mixte où des variables discrètes et continues sont mélangées. Nous avons d'autre part établi que, dans la démarche classique, qui part d'une structure statistique donnée, les équations du maximum de vraisemblance sont exactement les équations du PME associées à une information empirique. Les deux principes se renforcent alors mutuellement.

Nous avons poursuivi notre travail sur la relecture de la procédure de régression logistique par le PME. Nous avons montré que le PME permet d'envisager la procédure de régression logistique comme l'écriture naturelle d'une étude de régression où l'information apporté par un jeu de données (issu d'un échantillonnage de type quelconque) est résumée simplement par la loi empirique des variables explicatives, celle de la variable de groupe et par les moyennes empiriques des liens entre chaque variable explicative et la variable de groupe. Cette relecture évite d'avoir à justifier l'utilisation des équations issues de la maximisation d'une vraisemblance correspondant à un échantillonnage prospectif pour d'autres types de protocoles (dont nous avions prouvé qu'il n'est justifié qu'asymptotiquement) [108].

L'abord par le PME permet par ailleurs dans le cas très simple de la régression linéaire gaussienne d'intégrer la connnaissance *a priori* de la matrice variance des variables explicatives. La prise en compte de cette information est impossible par la méthode du maximum de vraisemblance. Le PME fournit ainsi un estimateur de variance plus faible que l'estimateur classique pour de très petits échantillons.

Enfin, une application de la modélisation par PME est envisagée dans le but de permettre le déplacement autonome d'un robot dans un environnement encombré d'obstacles, en se basant sur des connaissances préalables *a priori* et expérimentales (travail effectué en collaboration avec des informaticiens du laboratoire Leibniz). Cela permettra d'envisager le PME comme outil de modélisation en vue d'obtenir un modèle spécifiquement adapté dans des situations simples mais non standards. Un premier modèle où les histogrammes sont résumés par leurs modes a été construit et est en cours d'expérimentation.

3.2 Modèles hétéroscédastiques

Participants : Christian Lavergne, Catherine Trottier, Yann Vernaz

3.2.1 Modèles linéaires généralisés à effets aléatoires

Participants : Christian Lavergne, Catherine Trottier

L'incapacité des modèles linéaires classiques à permettre une analyse satisfaisante de certaines données comme les données binaires, a conduit à élargir l'ensemble des lois considéré et à définir les modèles linéaires généralisés (GLM). D'autre part, les modèles linéaires mixtes (L2M), dans lesquels les effets aléatoires sont venus compléter les effets fixes, ont aussi été introduits. De nombreux auteurs ont abordé l'analyse de ces modèles et plus particulièrement l'estimation des composantes de la variance. La combinaison de ces deux extensions des modèles linéaires classiques aboutit à la définition des modèles linéaires généralisés mixtes (GL2M). L'étude de ces modèles est la base du travail de thèse que C. Trottier a démarré en octobre 94.

Une nouvelle méthode d'estimation

Nous avons proposé une méthode d'estimation des composantes de la variance, dans le cadre des GL2M avec lien canonique, que nous mettons en parallèle avec la méthode décrite par Schall (1991). Cette méthode s'inspire à la fois de la théorie des GLM et des L2M ; elle présente deux étapes : la première dite "étape de linéarisation" s'inspire davantage de l'aspect GLM du GL2M, la seconde, dite "étape d'estimation", utilise le caractère aléatoire (L2M) du modèle linéarisé [106].

Extension de la méthode GAR

Gilmour, Anderson et Rae (1985) ont proposé une méthode d'estimation (GAR) pour des données binomiales dans un modèle avec lien probit. En 1993, Foulley et Im ont adapté la méthode GAR à des

données poissonniennes. Pour ces deux modélisations, C. Trottier a proposé une nouvelle lecture de la méthode en levant l'hypothèse d'homogénéité des variances des variables sous-jacentes. Elle propose ensuite une adaptation à des données de loi exponentielle et donne pour finir une formalisation qui permet d'unifier ces trois cas et d'envisager le cas de données binomiales dans un modèle avec lien logit (régression logistique)[113].

3.2.2 Modèles ARCH

Participants : Christian Lavergne, Yann Vernaz

La théorie des modèles ARCH (Auto-Régressifs Conditionnellement Hétéroscédastiques) introduite par Engel en 1982 peut à juste titre être considérée comme un des développements les plus prometteurs de la décennie pour modéliser le comportement des cours boursiers. Cette classe de modèles non linéaires, caractérisés par une variance conditionnelle, permet de déceler des périodes de volatilité plus faible ou plus forte au cours du temps. Ces modèles permettent aussi d'intégrer des propriétés observées empiriquement sur les séries financières : la dépendance quadratique entre deux observations, la forte sensibilité des variations sur les variations futures, les distributions à queues lourdes (leptokurtisme) des rentabilités.

L'étude de ces modèles est la base du travail de la thèse que Y. Vernaz a démarrée en octobre 96 et qui fait suite à un mémoire de DEA sur le sujet. Lors de ce premier travail, nous avons mis en œuvre une procédure itérative basée sur la méthode des moindres carrés (noté MCi) pour l'estimation des paramètres d'un modèle ARCH. Les résultats de simulations ont montré que cette procédure MCi était plus performante que l'algorithme BHHH (Bernt *et al.* 1974) classiquement utilisé pour calculer les estimations par pseudo-maximum de vraisemblance [116].

3.3 Algorithmes stochastiques

Participants : Gilles Celeux, Jean Diebolt

Nous avons poursuivi notre participation au groupe national MC.Cube, co-animé avec Dominique Cellier (université de Rouen) et C. Robert (CREST–ENSAE, Malakoff), sur le contrôle de la convergence vers la stationnarité des algorithmes stochastiques de type MCMC (*Monte Carlo Markov Chains*). Ce groupe s'est orienté vers l'étude approfondie du contrôle de la convergence de chaînes à espace d'états finis, ainsi qu'aux méthodes de couplage et de couplage arrière.

En collaboration avec Didier Chauveau (université de Marne-la-Vallée), nous avons entrepris de définir et d'expérimenter une procédure de contrôle utilisant des chaînes simulées en parallèle, afin de déterminer au moyen d'approximations empiriques un nombre minimal d'itérations au-delà duquel l'utilisation du théorème de la limite centrale puisse être considérée comme justifiée, en vue de construire des intervalles de confiance pour les résultats fournis par l'algorithme. Ce travail s'insère dans un projet de rédaction collectif de *Lecture Notes* regroupant les avancées du groupe MC.Cube.

Nous avions mis en évidence que la loi stationnaire des algorithmes stochastiques pour l'identification de mélanges de lois à k composants, présentaient $k!$ modes interchangeables, l'apparition de ces $k!$ modes étant due à l'indexation arbitraire des composants. Pratiquement cela peut rendre délicate l'exploitation des lois stationnaires des algorithmes stochastiques pour des mélanges imbriqués et pour lesquels l'information *a priori* est faible. Nous avons montré que les tentatives de résolution de ce problème d'identification par l'instauration de contraintes sur les paramètres du mélange sont décevantes. Elles induisent un fort biais causé par les accumulations artificielles de la loi stationnaire au bord des régions où les paramètres sont contraints de rester. À l'heure actuelle, nous cherchons des moyens d'éviter les renversements d'indiçage (*label switching*) sans construire une loi stationnaire déformée [100]. Ce travail se fait en liaison avec C. Posse (université du Minnesota). De plus, avec M. Bacha, nous explorons la capacité de l'algorithme BRM à éviter ces renversements d'indiçage.

Enfin, lors d'une visite à l'université de Washington (Seattle), G. Celeux a entrepris avec C. Posse et Guido Consonni (université de Pavie) une recherche sur l'extension au cadre bayésien hiérarchique de nos travaux sur l'analyse de mélanges gaussiens utilisant la décomposition spectrale des matrices variances des composants. L'intérêt de l'approche bayésienne hiérarchique réside dans un assouplissement des hypothèses faites sur les constituants géométriques (forme, volume et orientation) des classes. L'identification des modèles construits se fait par l'échantillonnage de Gibbs.

3.4 Modèles de fiabilité industrielle

Participants : Mostafa Bacha, Gilles Celeux, Jean Diebolt, Mhamed-Ali El Aroui, Edwige Idée, Joseph Ngatchou Wandji

Cette recherche s'effectue essentiellement dans le cadre de conventions d'étude et de recherche avec les groupes « Retour d'expériences » et « Fiabilité, Maintenance » de l'EDF-DER. On peut distinguer trois axes qui sont des modèles de durées de vie pour des systèmes fortement censurés, l'analyse bayésienne des défauts de cuves REP et l'estimation de queues de distributions. Un dénominateur commun à ces thèmes est qu'il concerne la modélisation d'événements rares.

3.4.1 Analyse de durées de vie de systèmes complexes

Participants : Mostafa Bacha, Gilles Celeux, Edwige Idée,

Dans le cadre de notre collaboration avec le groupe « Fiabilité, Maintenance » d'EDF, nous continuons la recherche sur la modélisation et l'estimation de durées de vie. Cette collaboration qui dure depuis plus de trois ans et a abouti à la thèse de M. Bacha [87], soutenue le 5 mars 1996, se prolonge sous la forme d'une recherche post-doctorale dont le but principal est le transfert des résultats de la thèse en codes de calcul directement utilisables par EDF.

Modèles de dépendance

Après avoir traité le cas des systèmes séries et des systèmes parallèles, nous nous sommes intéressés aux systèmes séries-parallèles par le modèle de dépendance de Gumbel. Nos investigations nous ont permis de constater la grande difficulté à estimer, dans des conditions réalistes, les paramètres des modèles de durées de vie de pareils systèmes. En effet, le nombre très limité d'observations disponibles et l'absence de données relatives à certains composants du système étudié font que la qualité des estimations est parfois médiocre car très sensible aux fluctuations des estimations des autres paramètres qui interviennent (c'est le cas pour l'estimation du paramètre de dépendance connaissant les autres paramètres du modèle). Dans nos études passées, nous avons privilégié le modèle de Gumbel qui est séduisant par sa simplicité, sa facilité d'interprétation et de simulation. Mais, il présente un défaut majeur, surtout dans un contexte industriel : l'introduction d'une dépendance entre les composants provoque une *amélioration* de la durée de vie du système. Cette limitation, partagé par de nombreux modèles, signifie que si X et Y désignent la durée de vie du couple, le modèle de dépendance induit l'inégalité

$$P(X > x, Y > y) \geq P(X > x)P(Y > y).$$

Cette propriété est irréaliste : les dépendances entre composants sont toujours perçues comme préjudiciables. Aussi, nous nous sommes intéressés dans le cadre du stage ENSIMAG de X. Horion [111] à la construction d'un modèle décrivant le comportement de deux composants montés en série en utilisant un paramètre permettant d'évoluer entre deux situations extrêmes : la première situation correspondant à une détérioration du meilleur composant jusqu'au niveau du moins bon, la deuxième correspondant à une amélioration du moins bon composant jusqu'au niveau du meilleur. Il est important de souligner que ce modèle ne mesure pas une liaison entre les lois mais un changement dans les lois marginales, un paramètre de liaison réglant l'amplitude de la dégradation ou de l'amélioration. Ce modèle, qui n'a de sens que si les paramètres des lois marginales sont connues avant la mise en fonctionnement en

série des composants, a été étudié pour des lois exponentielles pour lesquelless il donne des résultats intéressants et s'étend facilement à des lois de Weibull.

Modèles de Weibull fortement censurés

Un aboutissement important de la thèse de M. Bacha a été de donner les moyens pour choisir la méthode d'identification la plus fiable, en fonction du contexte, pour estimer les paramètres d'une loi de Weibull ou d'une loi de Poly-Weibull (associée à des composants montés en série). Les algorithmes en compétition sont les algorithmes EM, SEM et BRM. Ainsi, nous réalisons un guide détaillé pour choisir l'une de ces méthodes suivant le nombre de défaillances observées et le degré de censure. Ce guide inclut un questionnaire pour la construction des lois *a priori* lorsque la solution bayésienne est préconisée [87],[109].

Enfin, un stage ENSIMAG a porté sur le modèle d'Arrhenius, modèle classique de durée de vie prenant en compte l'environnement. Il s'avère qu'il revient à un simple modèle de mélange de lois exponentielles ou de Weibull et s'identifie simplement par l'algorithme EM [112].

3.4.2 Modélisation de caractéristiques de défauts

Participants : Gilles Celeux, Joseph Ngatchou Wandji

Il s'agit dans le cadre d'une convention d'étude et de recherche avec le groupe « Retour d'expériences » de EDF-DER de proposer et d'identifier un modèle probabiliste pour la taille et le nombre de défauts pouvant apparaître sur des cuves REP. Ce modèle doit tenir compte de l'incertitude sur les mesures de défauts et de la probabilité de détecter un défaut sachant sa taille. De plus, pour ce type de problème, les données de retour d'expérience sont extrêmement rares. Une méthodologie bayésienne a été développée. Les lois envisagées pour la modélisation bayésienne sont une loi de Weibull ou une loi log-normale pour la taille des défauts et une loi de Poisson pour leur nombre. Les lois *a priori* choisies sont des lois conjuguées classiques ou, pour le cas de la loi de Weibull, s'inspirent de celles utilisées dans [109]. L'identification du modèle assez complexe se fait par l'échantillonnage de Gibbs. Le but final est de fournir les éléments nécessaires pour la définition d'un cahier des charges précis et complet pour la réalisation d'un logiciel de détermination bayésienne des distributions de probabilité pour les défauts de cuves REP.

3.4.3 Modélisation et estimation de queues de distributions

Participants : Jean Diebolt, Mhamed-Ali El Aroui,

Nous étudions, dans le cadre d'une convention d'étude et de recherche avec le groupe « Retour d'expériences » de EDF-DER, le problème d'estimation des probabilités d'événements rares, ou queues de distribution. Plus précisément, si X est une variable aléatoire, le problème peut se résumer en l'estimation du quantile q_m défini par :

$$P(X > q_m) = \frac{1}{m} \quad m \text{ étant un réel positif} >> 1.$$

La première étape de cette étude consiste en une analyse critique et comparative (aussi bien théorique qu'expérimentale) des trois principales méthodes d'estimation des queues de distributions : l'utilisation de la théorie des valeurs extrêmes, la méthode des excès, et l'estimateur de Hill généralisé.

3.5 Statistique biomédicale

3.5.1 Modélisation de la croissance des thromboses

Participants : Jean-Luc Bosson, Gilles Celeux

Actuellement, pour éviter des embolies, toute thrombose décelée est traitée. Mais, grâce à l'amélioration du dépistage, on décèle des thromboses de plus en plus tôt. Il n'est pas sûr que le traitement systématique des thromboses décelées soit la meilleure solution, car nombre de petites thromboses se régulent d'elles-mêmes. Ainsi, un modèle pertinent sur la croissance des thromboses permettrait de définir une politique plus rationnelle de leur traitement. Pour construire un tel modèle, nous avons disposé d'une grande base de données (plus de 1000 patients d'origine hospitalière très diverse) dont on peut supposer qu'elle est représentative de tous les stades de développement possibles des thromboses veineuses. Partant de l'hypothèse que la croissance d'une thrombose dépendait de sa localisation, nous avons expérimenté des mélanges de lois exponentielles (suggérées par les histogrammes), dans le cadre du stage ENSIMAG de F. Carpentier [110]. Les algorithmes utilisées furent l'algorithme EM et l'algorithme SEM et donnèrent des résultats similaires. Finalement, c'est un modèle à trois composants -isolant une classe de très petites thromboses- qui s'avère le plus pertinent. Cependant, la qualité d'ajustement n'est pas excellente, et la recherche en utilisant des lois plus complexes que la loi exponentielle est envisageable. De plus, le caractère symptomatique ou asymptomatique des thromboses décelées mérite d'être contrôlé.

3.5.2 Le diagnostic des accidents vasculaires cérébraux au lit du malade

Participant : Claudine Robert

Cette recherche s'effectue en collaboration avec G. Besson (CHU de Grenoble). Le but est de rendre utilisables les futures thérapies pour le soin immédiat des accidents vasculaires cérébraux (AVC) en cas d'infarctus non hémorragiques (INH). Cette année, nous avons cherché à diagnostiquer les hémorragies ou à définir des cas-types d'hémorragie cérébrale. Seule la méthode d'induction par arbre a permis d'envisager la définition d'un syndrome, qui devra être validé (il faudra attendre au moins un an avant d'avoir le nombre de cas nécessaires à une telle validation).

Nous avons par ailleurs étudié les notions classiquement utilisées dans le champ de la médecine qui sont les syndromes, les critères diagnostiques et les scores (définitions, objectifs, méthodes de construction et de validation, aspects historiques). Le but était de montrer que tous ces concepts dérivent de la notion de score et donc que leur définition relève dans la plupart des cas de méthodes statistiques adaptées aux objectifs. L'ensemble de la recherche sur le diagnostic des accidents vasculaires cérébraux au lit du malade a fait l'objet de la thèse de G. Besson [88].

3.5.3 Analyse de la durée de séjour en gériatrie

Participants : Jean-Luc Bosson, Gilles Celeux, Claudine Robert

Le but est de proposer une description pertinente des durées de séjour en pédiatrie afin d'optimiser la gestion des ressources humaines et hôtelières et d'en déduire des modèles efficaces de prévision des lits nécessaires. La modélisation par un mélange de lois exponentielles des durées de séjour sera le modèle privilégié pour évaluer la classification administrative en trois classes (séjour court, moyen et long) et la classification en deux classes du Professeur Millard (pédiatre à Londres), qui propose un modèle analogue au nôtre ajusté par des moyens graphiques. Cette année, le travail a consisté à recueillir une base de données fiable et complète, et enrichie par l'enregistrement de variables cliniques et sociologiques, sur les durées de séjour en gériatrie au CHU de Grenoble. Le travail de modélisation proprement dite va se faire dans le cadre d'un stage de DEA, nous espérons le poursuivre par une thèse.

Signalons par ailleurs que nous avons renouvelé la demande de financement auprès de l'INSERM pour l'étude sur la durée de séjour hospitalière, pilotée par Georges Weil (CHU de Grenoble). Notre

précédente demande, dans le cadre d'un appel d'offres CNAM-INSERM avait été sélectionnée, mais suite à un désaccord entre ces deux organismes, l'appel d'offres n'avait pas été suivi d'effet.

3.6 Évaluation non paramétrique de modèles

3.6.1 Tests d'adéquation de modèles d'autorégression

Participants : Jean Diebolt, Joseph Ngatchou Wandji

Nous poursuivons notre exploration d'une procédure générale, reposant sur des processus empiriques (pondérés) des résidus et les théorèmes limites fonctionnels et principes d'invariance associés, pour tester selon une méthodologie de nature non paramétrique l'adéquation de modèles paramétriques de régression ou d'autorégression d'ordre un (y compris les modèles de type bilinéaire).

Cette année, nous avons fini d'explorer le cas particulier où l'on suppose connus les vrais paramètres du modèle à valider ([94], [95], [103]). Nous avons commencé à aborder le cas général où ces paramètres ne sont pas connus, et où l'on utilise, pour définir le processus empirique des résidus, des estimateurs de ces paramètres obtenus par la méthode des moindres carrés conditionnels. Les résultats limites conduisent alors à des processus gaussiens dont la fonction de covariance dépend du modèle considéré et des vraies valeurs inconnues des paramètres. Cela pose de nombreux problèmes que nous nous proposons d'analyser. En particulier, pour obtenir un test de type Kolmogorov-Smirnov à partir de ces processus empiriques des résidus, il faudra tabuler la loi du maximum $Z = \sup_{t \in J} |X(t)|$ sur un intervalle compact J de la valeur absolue de processus gaussiens $X(t)$, dans des cas où seuls sont connus des résultats asymptotiques sur $P\{Z \geq a\}$ lorsque $a \to \infty$, résultats insuffisamment précis pour fournir de bonnes approximations de $P\{Z \geq a\}$ pour la gamme des valeurs de a pour lesquels cette probabilité est située entre 0.10 et 0.01.

Une perspective en vue est de tester l'adéquation de modèles ARCH pour des séries chronologiques d'origine financière, en utilisant les procédures de test décrites ci-dessus (cf. 3.2.2).

3.6.2 Approximation du facteur de Bayes

Participants : Henri Bertholon, Gilles Celeux

La comparaison des modèles sur lesquels travaille IS2 peut se faire par des critères de choix de modèles qui sont en fait des approximations de facteurs de Bayes. H. Bertholon vient d'entamer une thèse sur ce sujet. Son premier travail consiste à construire un critère de choix entre un modèle exponentiel et un modèle de Weibull pour l'analyse de durées de vie. Cette recherche correspond à une préoccupation forte des fiabilistes.

4 Actions industrielles

4.1 Contrats EDF-DER en fiabilité

Participants : Mostafa Bacha, Gilles Celeux, Jean Diebolt, Mhamed-Ali El Aroui, Edwige Idée, Joseph Ngatchou Wandji

IS2 a signé cette année trois conventions d'étude et de recherche (CERD) avec l'EDF-DER. De plus, une grande partie du travail pour une convention signée en 1995 s'est déroulée cette année. Le contenu scientifique de ces conventions est décrit en 3.4.

Par ailleurs, G. Celeux est conseiller scientifique auprès de la société EUROPSTAT, en liaison avec l'ISDF, pour des études sur l'impact des données manquantes pour le retour d'expérience et la fusion de données d'origines diverses (retour d'expérience, avis d'expert,etc.).

4.2 Contact avec Péchiney

Participant : Christian Lavergne

Ce contact a eu lieu avec Aluminium-Pechiney pour l'étude du déséquilibre électrique dans les cuves à électrolyse chez Pechiney Aluval. Il s'agissait de proposer et construire des outils statistiques appliqués à ce problème industriel, outils liés à la régression multivariée. Ce travail a fait l'objet du mémoire de DEA de Francis de Véricourt [115].

5 Actions nationales et internationales

5.1 Actions nationales

J.-L. Bosson et C. Robert animent le séminaire de statistique médicale du CHU de Grenoble. Dans ce cadre, C. Robert a une activité importante de conseil statistique auprès de médecins (étude des moyens de comparaison de l'effet d'antihypertenseurs, étude de la modélisation du volume des lésions cérébrales provoquées chez le rat, étude des tests de tendance proposés par Armitage, étude visant à abolir l'usage de la notion de puissance a posteriori dans les publications méthodologiques en médecine).

J. Diebolt est l'un des animateurs du groupe national MC.Cube sur le contrôle de la convergence des chaînes de Markov de Monte Carlo. Il est membre du comité scientifique des XXIX^{e} Journées de Statistiques (ASU et SSF), qui se tiendront à Carcassonne (26–30 mai 1997). Il a été rapporteur des thèses de E. Gouno (université de Marne-la-Vallée) et de M. Bacha (université de Rouen).

G. Celeux a organisé la réunion n^{o}33 du groupe « Retour d'expérience » de l'ISDF. Il a été rapporteur des thèses de J. Pinto (université de Rennes 1) et C. Ambroise (UTC Compiègne).

Par ailleurs, IS2 participe régulièrement au séminaire de statistique du LMC-SMS à Grenoble. Dans ce cadre, ils ont invité J.-L. Foulley (INRA), D. Cellier et D. Fourdrinier (université de Rouen).

5.2 Actions internationales

G. Celeux a été invité un mois au département de statistique de l'université de Washington à Seattle. Dans ce cadre, il a participé au workshop « Model-based clustering and spatial point process » et il a mené des recherches en collaboration avec A. Raftery (université de Washington), C. Posse (université du Minnesota) et G. Consonni (université de Pavie). Il a été conférencier invité aux XVII^{e} Rencontres Franco-Belges de Statisticiens, consacrées à l'analyse statistique des modèles de durées de vie à Marne-la-Vallée en novembre 1996. De plus, il poursuit sa collaboration avec le LEAD de l'université de Lisbonne, et il dirige la thèse qu'I. Brito, assistante au département d'économie de l'université de Lisbonne a démarré sur la comparaison de méthodes d'analyse discriminante. Enfin, il codirige la thèse de G. Haenni (université de Genève) sur la modélisation statistique des durées de séjour hospitalier en cardiologie.

J. Diebolt a été conférencier invité aux XVI^{e} Rencontres Franco-Belges de Statisticiens, consacrées à la statistique des processus à Bruxelles en novembre 1995. Il est membre du comité scientifique des XVII^{e} Rencontres Franco-Belges de Statisticiens, consacrées à l'analyse statistique des modèles de durée de vie, Marne-la-Vallée, novembre 1996. Il poursuit la direction de la thèse de Jacques Zuber, Ecole Polytechnique Fédérale de Lausanne, qui porte sur les tests d'adéquation de modèles de régression linéaire généralisée (GLM).

6 Diffusion des résultats

6.1 Enseignement

G. Celeux enseigne la partie sur les algorithmes stochastiques dans l'option « systèmes markoviens » du DEA de mathématiques appliquées de Grenoble. Il enseigne les méthodes d'analyse discriminante dans les DEA de sciences cognitives et d'instrumentation biologique et médicale de Grenoble.

J. Diebolt enseigne le cours sur les tests non paramétriques dans le DEA de mathématiques appliquées de Grenoble. Il a donné un cours de 10 heures sur les méthodes MCMC dans le cadre de la « formation à la recherche » du CREST–ENSAE, Malakoff.

6.2 Organisation de colloques et de cours

G. Celeux a organisé un cours de trois jours sur l'analyse discriminante sur variables qualitatives à Villard-de-Lans en mars 1996.

7 Publications

Thèses

[87] M. BACHA, *Inférence statistique pour des modèles de durées de vie et applications*, thèse de doctorat, université de Rouen, mars 1996.

[88] G. BESSON, *Le diagnostic des accidents vasculaires cérébraux au lit du malade*, thèse de doctorat, université J. Fourier, Grenoble 1, novembre 1996.

Articles et chapitres de livre

[89] H. BENSMAIL, G. CELEUX, « Regularized Discriminant analysis through eigenvalue decomposition », *Journal of the American Statistical Association 91(4)*, 1996.

[90] G. BESSON, M. HOMMEL, S. BESSON, M. GUIGNIER, C. ROBERT, « Scoring systems for the differential diagnosis of ischemic and hemorrhagic stroke ? », *Stroke 27*, 1996, p. 337–338.

[91] G. CELEUX, D. CHAUVEAU, J. DIEBOLT, « Stochastic versions of the EM algorithm: an experimental study in the mixture case », *Journal of Statistical Computation and Simulation 55*, 1996.

[92] G. CELEUX, G. SOROMENHO, « An entropy criterion for assessing the number of clusters in a mixture model », *Journal of Classification 13(2)*, 1996.

[93] J. DIEBOLT, E. IP, « A Stochastic EM algorithm for approximate the maximum likelihood estimate », *in : Practical MCMC*, W. Gilks, S. Richardson, et D. Spiegelhalter (réd.), Chapman and Hall, 1996.

[94] J. DIEBOLT, J. NGATCHOU-WANDJI, « A nonparametric test for general first-order autoregressive models with a mixed term », *Notes aux Comptes Rendus de l'Académie des Sciences de Paris série I, 322*, 1996, p. 577–582.

[95] J. DIEBOLT, J. NGATCHOU-WANDJI, « A nonparametric test for general first-order autoregressive models », *Scandinavian Journal of Statistics 23(4)*, 1996.

[96] J. DIEBOLT, C. POSSE, « On the density of the maximum of smooth Gaussian processes », *Annals of Probability 24(4)*, 1996.

[97] M. EL-AROUI, C. LAVERGNE, « Generalized linear models in software reliability: parametric and semi-parametric approaches », *IEEE on Reliability 45*, 1996, p. 463–470.

Communications à des congrès, colloques, etc.

[98] M. BACHA, G. CELEUX, E. IDÉE, A. LANNOY, D. VASSEUR, «Estimation de modèles de durées de vie fortement censurés», *in : XVIIe Rencontre Franco-Belge de Statisticiens. Marne-la-Vallée*, 1996.

[99] M. BACHA, G. CELEUX, «BRM-IS : un algorithme d'estimation bayésienne pour modèles à données incomplètes», *in : XXVIIIe journées de statistique. Québec*, 1996.

[100] G. CELEUX, «Bayesian inference for mixture: the label switching problem», *in : Model-based clustering and spatial point process workshop, university of Washington*, p. 273–276, 1996.

[101] G. CELEUX, «Estimation of failure times involving Weibull distributions via stochastic algorithms», *in : XVIIe Rencontre Franco-Belge de Statisticiens. Marne-la-Vallée*, 1996. Conférence invitée.

[102] B. CRÉMILLEUX, C. ROBERT, «A pruning method for decision trees in uncertain domains: applications in medecine», *in : Proceedings of the workshop Intelligent Data Analysis in Medicine and Pharmacology, ECAI 96*, p. 15–20, 1996.

[103] J. DIEBOLT, J. NGATCHOU-WANDJI, «A nonparametric test for first order bilinear models», *in : Actes des XVI^e Rencontres Franco-Belges de Statisticiens (M. Hallin, edr)*, 1996.

[104] E. IDÉE, M. BACHA, G. CELEUX, A. LANNOY, D. VASSEUR, «Dependency modelling for parallel pumps», *in : 10th seminar on Rotating Machinery Reliability and Maintenance, European Safety Reliability and Data Association. Chamonix*, 1996.

[105] E. IDÉE, M. BACHA, G. CELEUX, A. LANNOY, D. VASSEUR, «A probabilistic model taking into account the dependence between failure times distributions of components connected in series», *in : Proceedings of the International Conference on Probabilistic Safety Assessment and Management. Crete, Greece*, 1996.

[106] C. LAVERGNE, C. TROTTIER, «Estimation dans les modèles linéaires généralisées mixtes», *in : XXVIIIe journées de statistique. Québec*, 1996.

[107] D. VASSEUR, M. BACHA, A. LANNOY, G. CELEUX, E. IDÉE, «Apport d'un modèle de dépendance et du retour d'expérience à la conception d'un système série», *in : Xe colloque national de Fiabilité & Maintenabilité. Saint-Malo*, 1996.

[108] V. VENDITTI, G. CELEUX, C. ROBERT, «Définition de la procédure de régression logistique par le principe de maximum d'entropie», *in : XXVIIIe journées de statistique. Québec*, 1996.

Rapports de recherche et publications internes

[109] M. BACHA, G. CELEUX, «Bayesian estimation of a Weibull distribution in a highly censored and small sample setting», *rapport de recherche n°2993*, Inria, 1996.

[110] F. CARPENTIER, *Modèle de mélange bi-modal pour les thromboses veineuses profondes*, Mémoire, Ensimag, 1996.

[111] X. HORION, *Caractérisation de modèles de dépendance entre composants industriels*, Mémoire, Ensimag, 1996.

[112] N. PEIRANI, *Influence de l'environnement sur les durées de survie*, Mémoire, Ensimag, 1996.

[113] C. TROTTIER, «Les GL2M : extension de la méthode GAR», *rapport de recherche n°????*, Inria, 1996.

[114] D. VASSEUR, A. LANNOY, M. BACHA, G. CELEUX, E. IDÉE, «A probabilistic model for dependent components in series system reliability», *rapport de recherche n°HP-28/96/024/A*, EDF-DER, 1996.

[115] F. VÉRICOURT, *Étude statistique du déséquilibre électrique des cuves d'électrolyse sur le site de Dunkerque*, Mémoire, DEA de mathématiques appliquées de Grenoble, 1996.

[116] Y. VERNAZ, *Estimation des paramètres dans les modèles ARCH*, Mémoire, DEA de mathématiques appliquées de Grenoble, 1996.

Divers

[117] M. BACHA, G. CELEUX, E. IDÉE, A. LANNOY, D. VASSEUR, «Modélisation de la dépendance de composants en parallèle», Rapport final de convention de recherche et développement Inria - EDF, 1996.

8 Abstract

The project IS2 of Inria Rhône-Alpes is concerned with statistical modelling. Emphasis is placed on incomplete data models. Its main areas of applications are biomedical statistic and failure time models. In 1996, IS2 developed activities statistical modelling through a maximum entropy principle, generalized linear models with random effects, stochastic algorithms, Bayesian statistical analysis of industrial failure times, competing risk models, tails of distributions, and medical applications (diagnosis of haemorragic infarct, models for thrombose's increase, and analysis of hospital length of stays via a mixture of exponential distributions).