Rapport INRIA 1994 — Programme 5

Modélisation statistique et applications biomédicales

Avant-projet SYSTOL

3 mai 1995

Avant-projet SYSTOL

Modélisation statistique et applications biomédicales

Localisation : *Rhône-Alpes*

Mots-clés : aide au diagnostic (1), algorithme stochastique (1), analyse de durées de vie (1), modèle linéaire généralisé (1), modélisation statistique (1), principe de maximum d'entropie (1), santé publique (1).

1 Composition de l'équipe

Responsable scientifique

Gilles Celeux, Directeur de recherche INRIA

Personnel Université

Edwige Idée, MC, université de Savoie
Christian Lavergne, MC, ENSIMAG/INPG
Claudine Robert, Professeur, université Joseph Fourier Grenoble 1

Chercheurs doctorant

Mostafa Bacha, boursier INRIA, université de Rouen
Catherine Trottier, allocataire MRT, INPG
Véronique Venditti, vacataire CHU de Grenoble, université Joseph Fourier Grenoble 1

Collaborateurs extérieurs

Jean Diebolt, CR CNRS URA 1321, Paris 6
George Weil, MC, université Joseph Fourier Grenoble 1

2 Présentation du projet

L'action SYSTOL a pour but d'effectuer des recherches en modélisation statistique en visant notamment des applications dans le domaine biomédical. SYSTOL, du fait de liens forts, est hébergé par le laboratoire d'informatique médicale TIMC (IMAG et institut A. Bonniot). Les thèmes de recherche considérés par SYSTOL sont les modèles linéaires généralisés, les modèles à structure cachée ou données manquantes et les algorithmes stochastiques. Nous concevons l'activité de recherche en statistique médicale sous deux aspects. Le premier est une activité de terrain pour laquelle avant d'atteindre des problèmes statistiques neufs, il s'agit de résoudre des problèmes réels. L'autre est une activité plus théorique, issue en général des études précédentes, et dont les retombées dépassent le cadre médical.

3 Action de recherche

3.1 Le principe de maximum d'entropie

Participants : Gilles Celeux, Claudine Robert, Véronique Venditti

Le principe de maximum d'entropie (PME) renverse la démarche de la statistique classique par le fait qu'elle amène à la construction des modèles à partir des observables (les statistiques exhaustives) alors que dans l'approche traditionnelle les statistiques exhaustives sont déduites du modèle. L'intérêt de ce point de vue est double. D'une part, il permet une relecture de modèles classiques qui met en évidence les observables utilisés. D'autre part, il permet d'envisager la modélisation pour des observables complexes associés par exemple à des graphes de dépendance définis par les médecins. Cette année, nous avons précisé le formalisme du calcul des lois continues par le PME. Nous avons analysé des situations d'analyse discriminante par le PME. Cette recherche nous a notamment permis de montrer que la régression logistique était un modèle optimal, au sens du PME, dans de nombreuses situations. De plus, nous avons précisé les liens entre le PME et l'estimation du maximum de vraisemblance du modèle associé. Ce thème est l'axe directeur de la thèse que prépare Véronique Venditti.

3.2 Analyse de durées de vie de systèmes complexes

Participants : Mostafa Bacha, Gilles Celeux, Edwige Idée

Cette recherche s'effectue dans le cadre d'un contrat avec le groupe ≪Retour d'expériences≫ de l'EDF. Elle constitue l'axe de la thèse que prépare Mostafa Bacha. D'une part, nous avons proposé une solution bayésienne complète pour l'estimation des paramètres de forme et d'échelle d'une loi de Weibull à partir de durées de vie censurées. (La loi de Weibull est une loi de référence en analyse de durées de vie de matériels industriels car elle permet de modéliser aussi bien les défauts de jeunesse que le viellissement précoce.) Nous levons la difficulté de non appartenance de la loi de Weibull à la famille exponentielle de distributions, par l'usage de l'algorithme sc WLB-SIR de Newton et Raftery. De plus nous proposons une version plus naturelle de cet algorithme où la phase de tirages par le bootstrap de pseudo échantillons des paramètres est remplacé par des tirages des données censurées selon un principe d'attribution aléatoire. Nous comptons étendre cette approche au cas de systèmes constitués de composants indépendants montés en série et dont la durée de vie est régie par des lois de Weibull. Pour ce problème, nous avons cette année considéré l'estimation du maximum de vraisemblance des paramètres des lois par l'algorithme EM (Expectation-Maximisation). Les résultats sont satisfaisants, mais ne prennent pas en compte les informations a priori qui sont importantes en ce domaine.

Par ailleurs, nous avons proposé un modèle qui permet de prendre en compte la dépendance entre des composants montés en série. Dans le cas de deux composants X et Y en série suivant des lois de Weibull dont les paramètres de forme et d'échelle sont désignés par les lettres β et η, le modèle est caractérisé par la fonction de survie

$$R_{XY}(x,y) = \exp\left[-\left((\frac{x}{\eta_1})^{\frac{\beta_1}{\delta}} + (\frac{y}{\eta_2})^{\frac{\beta_2}{\delta}}\right)^{\delta}\right], \quad \delta \in]0,1].$$

et par prolongement par continuité pour $\delta = 0$, on pose

$$R_{XY}(x,y) = \mathrm{P}(X > x, Y > y) = \exp\left[-\max\left((\frac{x}{\eta_1})^{\beta_1}, (\frac{y}{\eta_2})^{\beta_2}\right)\right].$$

Le paramètre δ mesure la liaison entre les deux composants. En particulier, $\delta = 1$ correspond à l'indépendance des variables X et Y, alors

que $\delta = 0$ correspond à une liaison parfaite. Les paramètres β et η étant estimés par ailleurs, nous avons construit et testé différentes méthodes d'estimation de δ dont la meilleure s'avère être la résolution itérative des équations de vraisemblance.

3.3 Algorithmes stochastiques

Participants : Gilles Celeux, Jean Diebolt

Cette recherche s'effectue en collaboration Didier Chauveau (université de Marne la Vallée). Cette année, nous avons précisé d'un point de vue théorique les liens existant entre les versions stochastiques de l'algorithme EM et l'inférence bayésienne non informative. Ces liens ont notamment été explicités pour l'analyse de durées de vie censurées suivant des lois exponentielles. Par ailleurs, nous avons mis en évidence que pour l'analyse de mélanges de lois à k composants, les algorithmes stochastiques fondés sur la simulation de chaînes de Markov produisaient des lois stationnaires ayant $k!$ modes intrinsèques, ce qui induit des difficultés d'analyse réelles et jusque là négligées. Enfin, signalons que, depuis octobre, nous animons en collaboration avec Bernard Ycart (projet MAI de l'IMAG) un groupe de travail sur l'évaluation de la convergence des algorithmes stochastiques fondés sur la simulation de chaînes de Markov. Dans ce groupe, nous espérons faire collaborer des préoccupations pratiques et théoriques.

3.4 Modèles linéaires généralisés

Participants : Christian Lavergne, Catherine Trottier

Dans le cadre d'une convention avec le centre de recherche de Péchiney (Voreppe), nous avons proposé une présentation unifiée des modèles d'analyse de variance à 2 facteurs aléatoires sur plans équilibrés. Notre présentation met en lumière l'équivalence des différents estimateurs des modèles dans ce cas (estimateurs ANOVA, MINQUE et REML). Ce travail, effectué en collaboration avec Gilles Celeux, a fait l'objet du stage de DEA de Moulaye Hamza (LMC) et a donné lieu à l'élaboration d'un formulaire pour Péchiney des estimateurs et tests dans tous les cas possibles pour le type de modèle considéré.

Nous avons introduit une famille d'estimateurs des composantes de la variance dans un modèle linéaire mixte à k composantes de la variance utilisant une procédure de rééchantillonnage de type Jackknife sur des formes quadratiques. La famille introduite par une méthode de moindres carrés généralisés a la propriété remarquable d'inclure la famille des estimateurs MINQUE. On est donc amené à envisager une amélioration naturelle des estimateurs des composantes de la variance dans un modèle linéaire mixte par une procédure de maximum de vraisemblance.

Signalons enfin que la thèse de Catherine Trottier, entamée début octobre, concerne l'étude des modèles linéaires généralisés (dont la fonction de lien n'est pas linéaire) et qui comportent de plus des composantes à effets aléatoires. L'estimation de ce type de modèles sera une question centrale de sa thèse. Cette recherche pourrait se fonder sur une interprétation de ces modèles comme des modèles à données manquantes et utiliser des algorithmes stochastiques.

3.5 Statistique biomédicale

3.5.1 Analyse de la durée de séjour hospitalière

Participants : Gilles Celeux, Claudine Robert, Georges Weil

Il s'agit de construire un nouvel indicateur définissant le séjour hospitalier à partir du regroupement des séjours relatifs à une même pathologie. Cet indicateur sera fondé sur des informations médicales et administratives avec comme objectif de pouvoir se passser des informations médicales grâce aux résultats que fournira une analyse statistique des données administratives. Les enjeux sont élevés puisqu'un tel indicateur pourrait jouer un rôle important dans la description statistique des activités hospitalières. Cette année, nous avons effectué une analyse exploratoire des données administratives du CHU de Grenoble pour 1992 dans le cadre d'un stage de deux étudiants du département statistique de l'IUT de Grenoble. Cette action a fait l'objet d'une demande de financement auprès de la CNAM (demande CNAM-INSERM).

3.5.2 Facteurs pronostics pour le cancer du sein

Participant : Véronique Venditti

Véronique Venditti a constitué une base de données sur le cancer du sein contenant des facteurs classiques relatifs à la tumeur et des facteurs cytologiques. À l'heure actuelle, les analyses statistiques effectuées sont des régressions selon le modèle à hasard proportionnel de Cox. Ces premières analyses montrent que certains facteurs cytologiques pourraient avoir un pouvoir pronostic intéressant.

3.5.3 Analyse statistique de translocations réciproques

Participant : Christian Lavergne

Ce travail s'est fait en collaboration avec Christine Cans et Olivier Cohen (CHU de Grenoble). Son objet est l'étude statistique d'une anomalie de la structure des chromosomes, résultat d'un échange segmentaire entre deux chromosomes non homologues. Devant un parent porteur d'une translocation réciproque, les généticiens souhaitent disposer d'une estimation du risque de ≪ségrégation non alternée≫ conduisant à un enfant atteint de maladie génétique. Ce risque est appréhendé par une modélisation logistique de type modèle additif généralisé.

3.5.4 Les essais d'équivalence

Participant : Claudine Robert

Cette recherche s'est effectuée en collaboration avec Jean-Luc Bosson (TIMC-IMAG et CHU de Grenoble). Partant de l'étude théorique concernant la bioéquivalence pharmaceutique et de situations pratiques où l'on cherche à voir si la prescription de deux médicaments donne les mêmes résultats, nous avons défini ce qu'est un essai d'équivalence, par opposition aux essais d'efficacité, dans le cas de grands échantillons. Cette étude a par ailleurs incité Claudine Robert et Gilles Celeux à réfléchir sur le bien-fondé de l'analyse bayésienne en statistique biomédicale.

3.5.5 Le soin immédiat des accidents vasculaires cérébraux

Participant : Claudine Robert

Cette recherche s'effectue en collaboration avec Gérard 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). Les données cliniques en cas d'AVC avaient jusqu'à présent été jugées non pertinentes pour le diagnostic de la cause de l'AVC (thrombose ou hémorragie) car on ne pouvait disposer du diagnostic que par autopsie, c'est-à-dire pour un échantillon biaisé. Avec le scanner, on peut avoir ce diagnostic pour un échantillon non biaisé. Nous avons donc étudié un tel échantillon qui semble permettre la détection immédiate d'INH d'environ 35 % des cas d'AVC (environ 45 % des cas d'infarctus cérébraux). Ce diagnostic est en cours de validation. Il est important car il semble que bientôt, on disposera de thérapies immédiates permettant d'éviter des handicaps majeurs dus à l'INH.

4 Actions industrielles

4.1 Contrat EDF-Retour d'expérience : analyse de durées de vie

Participants : Mostafa Bacha, Gilles Celeux, Edwige Idée

Ce contrat portant sur l'étude de la dépendance des composants d'un système monté en série a donné lieu à une recherche décrite dans le paragraphe 3.2. Ce contrat va trouver son prolongement dans un nouveau contrat entre l'EDF et l'action SYSTOL qui portera sur l'identification de systèmes plus complexes comportant à la fois des composants montés en série et d'autres en parallèle.

4.2 Contrat Péchiney : analyse de variance

Participants : Gilles Celeux, Christian Lavergne

Ce contrat a consisté à rédiger un document de synthèse pour Péchiney sur les estimations et les tests valides à partir d'un modèle d'analyse de variance à 2 facteurs aléatoires sur plan équilibré (cf. paragraphe 3.4).

5 Actions nationales et internationales

5.1 Actions nationales

Claudine Robert anime le séminaire de statistique du CHU de Grenoble. Gilles Celeux anime, conjointement avec Bernard Ycart (projet MAI de l'IMAG), un groupe de travail sur l'évaluation de la convergence des algorithmes stochastiques.

5.2 Actions internationales

Gilles Celeux a été invité 6 semaines au département de statistique de l'université de Washington à Seattle. Celà lui a essentiellement permis de finaliser la recherche, entreprise lors du séjour d'Adrian Raftery (université de Washington) l'an dernier à l'INRIA Rocquencourt, sur l'exploitation de la décomposition spectrale d'une matrice de variance en discrimination et en classification bayésienne.

Gilles Celeux poursuit sa collaboration avec le LEAD de l'université de Lisbonne. Cette année, il s'est rendu 2 fois à Lisbonne, pour la thèse de Gilda Soromenho qu'il a dirigée et en tant que conférencier invité des 2^{e} journées d'analyse des données de la société portugaise de classification.

6 Diffusion des résultats

6.1 Enseignement

Gilles Celeux enseigne la partie sur les algorithmes stochastiques dans l'option ≪systèmes markoviens≫ du DEA de mathématiques appliquées de Grenoble. Claudine Robert enseigne l'analyse des données dans le DEA de GBM de l'université Lyon-I.

6.2 Organisation de colloques et de cours

Christian Lavergne est l'un des organisateurs des XV^{e} rencontres franco-belges de statistique qui se sont tenues en novembre à Villard de Lans et dont le thème était les ondelettes en statistique.

7 Publications

Livres et monographies

[1] G. Celeux, J.-P. Nakache, *Analyse discriminante sur variables qualitatives*, Polytechnica, 1994.

Articles et chapitres de livre

[2] C. Cans, C. Lavergne, «De la régression logistique vers un modèle additif généralisé : un exemple d'application», *Revue de Statistique Appliquée*, 1994, À paraître.

[3] G. Celeux, G. Govaert, «Gaussian parsimonious clustering models», *Pattern Recognition*, 1994, À paraître.

[4] G. Celeux, G. Soromenho, «An entropy criterion for assessing the number of clusters in a mixture model», *Journal of Classification*, 1994, À paraître.

[5] 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, 1994, À paraître.

[6] E. Idée, «Estimation des $k+1$ paramètres de la loi d'un système : formé de k composants en série», *Publications du Centre de Recherches en mathématiques pures de Neuchatel, SÉRIE II*, 1994, À paraître.

Communications à des congrès, colloques, etc.

[7] A. Antoniadis, C. Lavergne, «Modèles hétérocédastiques et ondelettes», *in : XV^{mes} rencontres Franco-Belges de Statistique*, Villard de Lans, 1994.

[8] M. Bacha, G. Celeux, J. Diebolt, E. Idée, «Estimating failure time distribution from censored systems arranged in series», *in : New approaches in Classification and Data Analysis*, E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, B. Burtschy (réd.), Springer-Verlag, p. 533–538, 1994.

[9] M. Bacha, G. Celeux, E. Idée, A. Lannoy, D. Vassseur, «Estimation bayésienne des paramètres d'une loi de Weibull», *in : Actes du congrès Qualité et sûreté de fonctionnement*, p. 112–118, Compiègne, 1994.

[10] M. Bacha, G. Celeux, «Discussion of a paper by Newton and Raftery», *in : Journal of the Royal Statistical Society, B*, 56, 1994.

[11] M. Bacha, «Estimation of the parameters of infimum of Weibull distributed failure time distributions», *in : COMPSTAT 94*, Springer-Verlag, p. 209–214, Vienne, 1994.

[12] C. Cans, O. Cohen, C. Lavergne, «Assessment of the risk of unbalanced offspring and prenatal diagnosis strategy in translocations», *in : XXVIth annual meeting of European Society of Human Genetics*, 1994.

[13] G. Celeux, G. Govaert, «Fuzzy clustering and mixture models», *in : COMPSTAT 94*, Springer-Verlag, p. 154–159, Vienne, 1994.

[14] G. Celeux, «L'algorithme SEM comme approximation de l'algorithme d'augmentation de données», *in : XXVI^{es} journées de statistique*, Neuchatel, 1994.

[15] G. Celeux, «Quelques aspects de la régression logistique et un principe de maximum d'entropie», *in : 2^{me} journées d'analyse des données de la société portugaise de classification*, Lisbonne, 1994. Conférencier invité.

[16] G. Celeux, «A Stochastic EM algorithm as an approximation of Data Augmentation for non informative Bayesian inference», *in : International worhhop on highly structured stochastic systems*, Cortona, 1994. Conférencier invité.

[17] E. Idé, «Théorie de la fiabilité et utilisation du logiciel FIABAYES», *in : Conférence sur l'approche probabiliste de sureté*, EDF, 1994.

Rapports de recherche et publications internes

[18] M. Bacha, G. Celeux, E. Idée, A. Lannoy, D. Vassseur, «Estimation de la dépendances de lois de Weibull montées en série», *Rapport final de convention inria-edf*, INRIA-EDF, 1994.

[19] H. Bensmail, G. Celeux, A. Raftery, C. Robert, «Bayesian clustering inference», *rapport de recherche*, Departement of Statistics, University of Washington, 1994, À paraître.

[20] H. Bensmail, G. Celeux, «Regularized Discriminant analysis through eigenvalue decomposition», *rapport de recherche n°278*, Departement of Statistics, University of Washington, 1994.

[21] J.-L. Bosson, C. Robert, «The therapeutic Δ equivalence. The case of large samples», *rapport de recherche*, INRIA, 1994, À paraître.

[22] C. Lavergne, G. Celeux, «Les modèles d'ana;yse de la variance à deux facteurs avec effets aléatoires sur plan équilibré : une présentation exhaustive», *Rapport final de convention inria-péchiney*, INRIA-Péchiney, 1994.

8 Abstract

The action SYSTOL of INRIA Rhône-Alpes is concerned with statistical modelling. Its main area of applications is biomedical statistic. In 1994, SYSTOL developed activities on models arising from a maximum entropy principle, generalized linear models with random effects, stochastic algorithms and industrial failure lifetime analysis. Medical applications concerned using logistic regression for reciprocal translocation, assessing cytologic factors for the pronostic of the breast cancer, diagnosis of haemorragic infarct, therapeutic Δ equivalence, and analyzing hospital length of stays.

Table des matières