Rapport sur le manuscrit de thèse de

Etienne DANTAN

**"Modèles conjoints pour
données longitudinales et données de survies incomplètes
appliqués à l'étude du vieillissement cognitif"**

L'objectif de ce travail de thèse est de présenter des méthodes statistiques pour l'analyse de données longitudinales ciblées sur l'épidémiologie du vieillissement cognitif.

Le chapitre d'introduction pose clairement la problématique de l'étude du vieillissement cognitif et des enjeux de la méthodologie statistique à développer en particulier lorsque l'on s'intéresse à la survenue d'une démence. Il se termine par l'organisation du mémoire. L'intérêt de ce chapitre est évident puisqu'il a le mérite d'énoncer en quelques pages les mécanismes de la cognition et permet ainsi de comprendre les questionnements statistiques développés dans ce travail de thèse. On a toutefois du mal à identifier les problèmes laissés en suspend et la nature des jeux de données qui seront étudiés par la suite.

Le chapitre 2 présente un état de l'art détaillé il est vrai, mais qui donne le sentiment d'un simple catalogue. Bien que bien écrit, on ne comprend pas le fil directeur de ce chapitre. Il est décrit de multiples modèles sans savoir vraiment lesquels vont être repris dans la suite/ On voit aussi par exemple des phrase du type : "'Une solution envisageable est ... . Dans de nombreux travaux, l'algorithme classiquement utilisé est ..." et on se retrouve en fin de ce chapitre avec une description de méthodes d'extimation qui ne sont pas véritablement mises en oeuvre.

Le cœur de la thèse est composée des chapitres 3 et 4.

Chacun de ces 2 chapitres est en fait basé sur un article publié :

Pattern Mixture Models and Latent Class Models for the Analysis of Multivariate LOngitudinal Data with Informative Dropouts publié dans The International Journal of Biostatistics pour ce qui est du chapitre 3.

Joint model with latent state for longitudinal and multi-state data publié (article soumis) pour le chapitre 4. Cet article est complété de simulations et d'une application à la cohorte Paquid.

Chapitre 3 : Comme le suggère le titre de l'article, il s'agit ici de comparer les 2 approches compte tenu de l'hypothèse faite sur les données manquantes : sont elles ou pas informatives ? Cette comparaison est faite sur un modèle longitudinal à processus latent (Proust-Lima 2006) et nous apprends comme on pouvait s'y attendre que l'interprétation induite des estimations des paramètres est sensiblement différente dans les 2 approches. Ce chapitre 3 est réduit à cet article publié en janvier 2008 et on peut s'étonner que dans la rédaction de ce manuscrit de thèse (2 ans après l'écriture de l'article) il n'y ait aucun autre commentaire ou extension naturelle.

Chapitre 4 : Clairement le reproche précédent ne s'applique pas ici. L'article est ici complété de 35 pages de simulations bien détaillées et l'application est très poussée. Sur le fond l'article de ce chapitre reprend un modèle développé par Jacqmin-Gadda en 2006 pour l'étude conjointe de l'accélération du déclin cognitif et de la survenue de la démence. Ce modèle est étendu à un modèle à état

latent ; cette extension est un apport important à ce qui se faisait précédement. La discussion de cette approche (limites et perspectives) amène un eclairsissement intéressant. Seule la discussion sur l'aspect numérique semble un peu sommaire et on reste un peu sur une interrogation : pourquoi être rester sur Newton-Raphson alors que comme le dit l'auter des algorithmes de type MCMC autoriserait une plus grande marge de manoeuvre.

Dans le chapitre 3 est proposé un modèle non linéaire à processus latent. La première partie de ce chapitre est constituée d'un article publié dans *Biometrics* et la 2e partie est construite autour de simulations sur le modèle proposé.

Dans le chapitre 4 est proposé un modèle non linéaire à classes latentes (autrement dit un modèle de mélange fini) pour l'analyse conjointe de plusieurs marqueurs quantitatifs et d'une variable binaire. Ce chapitre est en fait un article publié dans *Statistics in Medecine*. Ces 2 articles publiés témoignent bien de l'opportunité et de la qualité de ce travail.

Dans le chapitre 5 est proposé un modèle non linéaire à classes latentes pour l'analyse conjointe de plusieurs marqueurs quantitatifs et d'un temps d'événement. Ce modèle est proposé pour pallier aux limitations du modèle précédent.

Ces 3 chapitres sont le cœur de cette thèse. Ils sont indéniablement dans la problématique posée, apportent des avancées significatives au niveau de la modélisation, et sont bien dans une logique de "progression" et "sophistication" dans la modélisation de l'évolution cognitive. En effet lorsque l'on regarde de plus prêt le modèle du chapitre 5, il comporte conjointement : une modélisation de type modèle mixte pour tenir compte de l'aspect longitudinal des données, un modèle de survie à risque proportionnel (avec des individus censurés à gauche pour tenir compte de la fin de la fenêtre d'observation) ; et sur ces 2 modèles conjoints est greffé le modèle de mélange fini. Et pour finir, Cécile Proust enrichit ce modèle par la modélisation du risque par des fonctions splines. Cette richesse semble "justifiée" compte tenu de la problématique, mais l'aspect numérique est laissé sous silence. On se doute bien que l'on va finir par être confronté par des limitations algorithmiques. Hormis la phrase p133 "*Le modèle* (il s'agit du modèle de Weibull) *utilisant des splines était néanmoins trop compliqué et l'estimation n'a pas abouti au bout de 500 itérations*" on ne retrouve pour ainsi dire pas de discussion sur le sujet. Est-ce à dire qu'il n'y a pas de difficultés numériques ? Bien entendu, l'ensemble des simulations construit au cours des chapitres améne déjà des réponses ; elles sont suffisamment riches et les résultats convaincants. Il est tout aussi clair que l'investissement sur le coté informatique du travail a été très important ; une façon de ne pas le laisser sous silence serait par exemple de donner une adresse http pour accéder à l'ensemble des programmes. Ceci aurait aussi pour conséquence une large diffusion.

En ce qui concerne l'application sur le jeu de données PAQUID, mené tout au long des chapitres 3, 4 et 5, on y voit une grande richesse de développements, rediscutés dans le chapitre 6 aussi bien en termes de limitation que de perspectives. Il est clair que l'étude de l'épidémiologie du vieillissement cognitif n'est pas ici un prétexte. Sûrement que le milieu médical a su ou saura en tirer profit. Une question que je me pose : serait-il pertinent de faire une décomposition du gain de la vraisemblance pour chaque structure cachée ? Évidemment cette décomposition n'est pas orthogonale mais permet dans certaines situations de comprendre l'apport dans la modélisation de chacune des structures. *La partie 3 est celle qui va se rapprocher le plus de ce travail de thèse : la modélisation conjointe de marqueurs longitudinaux et*

d'un événement. On y trouve la description des modèles à effets aléatoires partagés et des modèles conjoints à classes latentes. Une sommaire comparaison de ces 2 types de modèles est proposée avec des arguments de modélisation, d'interprétation et algorithmiques. C'est un point intéressant et on aurait aimé avoir une étude plus approfondie en particulier avec des exemples de données pour bien appréhender les limitations de chaque modélisation.

Dans le chapitre 3 est proposé un modèle non linéaire à processus latent. La première partie de ce chapitre est constituée d'un article publié dans Biometrics et la 2e partie est construite autour de simulations sur le modèle proposé.

Dans le chapitre 4 est proposé un modèle non linéaire à classes latentes (autrement dit un modèle de mélange fini) pour l'analyse conjointe de plusieurs marqueurs quantitatifs et d'une variable binaire. Ce chapitre est en fait un article publié dans Statistics in Medecine. Ces 2 articles publiés témoignent bien de l'opportunité et de la qualité de ce travail.

Dans le chapitre 5 est proposé un modèle non linéaire à classes latentes pour l'analyse conjointe de plusieurs marqueurs quantitatifs et d'un temps d'événement. Ce modèle est proposé pour pallier aux limitations du modèle précédent.

Ces 3 chapitres sont le cœur de cette thèse. Ils sont indéniablement dans la problématique posée, apportent des avancées significatives au niveau de la modélisation, et sont bien dans une logique de "progression" et "sophistication" dans la modélisation de l'évolution cognitive. En effet lorsque l'on regarde de plus prêt le modèle du chapitre 5, il comporte conjointement : une modélisation de type modèle mixte pour tenir compte de l'aspect longitudinal des données, un modèle de survie à risque proportionnel (avec des individus censurés à gauche pour tenir compte de la fin de la fenêtre d'observation) ; et sur ces 2 modèles conjoints est greffé le modèle de mélange fini. Et pour finir, Cécile Proust enrichit ce modèle par la modélisation du risque par des fonctions splines. Cette richesse semble "justifiée" compte tenu de la problématique, mais l'aspect numérique est laissé sous silence. On se doute bien que l'on va finir par être confronté par des limitations algorithmiques. Hormis la phrase p133 "Le modèle (il s'agit du modèle de Weibull) utilisant des splines était néanmoins trop compliqué et l'estimation n'a pas abouti au bout de 500 itérations" on ne retrouve pour ainsi dire pas de discussion sur le sujet. Est-ce à dire qu'il n'y a pas de difficultés numériques ? Bien entendu, l'ensemble des simulations construit au cours des chapitres amène déjà des réponses ; elles sont suffisamment riches et les résultats convaincants. Il est tout aussi clair que l'investissement sur le coté informatique du travail a été très important ; une façon de ne pas le laisser sous silence serait par exemple de donner une adresse http pour accéder à l'ensemble des programmes. Ceci aurait aussi pour conséquence une large diffusion.

En ce qui concerne l'application sur le jeu de données PAQUID, mené tout au long des chapitres 3, 4 et 5, on y voit une grande richesse de développements, rediscutés dans le chapitre 6 aussi bien en termes de limitation que de perspectives. Il est clair que l'étude de l'épidémiologie du vieillissement cognitif n'est pas ici un prétexte. Sûrement que le milieu médical a su ou saura en tirer profit. Une question que je me pose : serait-il pertinent de faire une décomposition du gain de la vraisemblance pour chaque structure cachée ? Évidemment cette décomposition n'est pas orthogonale mais permet dans certaines situations de comprendre l'apport dans la modélisation de chacune des structures.

En conclusion je dirai que l'objectif annoncé de ce travail est parfaitement atteint.

C'est à la fois un travail de statistique appliquée moderne qui se situe dans le contexte général des modèles mixtes, modèles de survie, modèles de mélanges, et un travail d'épidémiologie.

L'apport en statistique est indéniable et la diffusion de ce travail en dehors de la communauté "santé publique" sera un vrai gain pour de nombreux autres domaines de recherche.

Je donne donc un avis favorable à ce que cette thèse soit soutenue devant l'Université Victor Segalen, Bordeaux II.

Fait à Montpellier le 5 novembre 2009

Christian Lavergne
Professeur à l'Université Paul Valéry - Montpellier III
email : Christian.Lavergne@univ-montp3.fr