Les modèles de Markov cachés à effets mixtes

Maud Delattre

Département de Mathématiques
Bâtiment 425
Faculté des Sciences d'Orsay
Université Paris-Sud 11
F-91405 Orsay Cedex

Résumé court

Les modèles de Markov cachés à effets mixtes sont des modèles très récents. Ils se définissent comme l'extension des modèles de Markov cachés classiques aux études de population. Dans les cas où l'on peut dissocier plusieurs états dans une maladie, ces nouveaux modèles sont particulièrement adaptés à l'analyse de données longitudinales recueillies lors d'essais cliniques. Toutefois, l'estimation pour ces modèles est une problématique complexe. En particulier, les modèles de Markov cachés à effets mixtes présentent une structure hautement non linéaire et certaines données ne sont pas observées, ce qui complique grandement l'expression de la vraisemblance et sa maximisation. Nous proposons une méthodologie complète d'apprentissage de ces modèles en trois étapes. Un algorithme EM stochastique combiné à l'algorithme de Baum-Welch permettra d'abord d'estimer les paramètres de population de nos modèles. Ensuite, nous estimons les paramètres des modèles de Markov cachés individuels par maximisation de leur distribution a posteriori. Enfin, les séquences d'états les plus probables au vu des données sont obtenues par l'algorithme de Viterbi. Des études de Monte Carlo ainsi qu'une première application à des données d'épilepsie renvoient des résultats encourageants du point de vue de la qualité des estimateurs obtenus.

Summary

Mixed-effects hidden Markov models have been newly defined in the statistical literature as an extension of hidden Markov models for dealing with population studies. The notion of mixed hidden Markov models is particularly relevant for modeling longitudinal data collected during clinical trials, especially when distinct disease stages can be considered. However, parameter estimation in mixed hidden Markov models is complex, especially due to their non-linear structure and the presence of hidden data. Expressing and maximizing likelihood in these models is thus complicated. For that purpose, we propose a new procedure, that can be briefly described as follows. First, we suggest estimating the population parameters with a stochastic approximation EM algorithm coupled with the Baum-Welch algorihtm. Then, for dealing with the individuals, we estimate each

set of individual parameters with the MAP (Maximum A Posteriori) of the parameter distributions. Finally, the hidden state sequences are decoded using the Viterbi algorithm. Monte Carlo simulations and a first application to epilepsy data show very encouraging results.

Mots clés

Biostatistique, processus, modèles de Markov cachés, modèles mixtes, algorithme SAEM, algorithme de Baum-Welch, maximum a posteriori, algorithme de Viterbi.

Résumé long

Comme l'ont déjà montré certains travaux portant sur la migraine ([5]) ou l'épilepsie ([1]), les modèles de Markov cachés (*Hidden Markov Models, HMM*) sont des candidats intéressants pour la modélisation des symptômes liés à certaines maladies chroniques. En effet, ces modèles supposent que la progression de la maladie peut s'interpréter à travers différents stades ou degrés de gravité, qui conditionneraient l'intensité des manifestations symptômatiques. Ainsi, chaque malade séjournerait alternativement dans chaque état pendant des périodes plus ou moins longues. Plus précisément, on considère dans les *HMM* que les stades évolutifs de la maladie correspondent aux états d'une chaîne de Markov non observée, les seules informations individuelles accessibles étant les valeurs de marqueurs biologiques en des temps donnés du suivi du patient, marqueur dont la distribution de probabilité est supposée changer selon la nature de l'état. Par exemple, comme dans les travaux d'Albert (1991) sur quelques patients épileptiques, il paraît raisonnable d'interpréter l'évolution du nombre quotidien de crises d'épilepsie en dissociant deux états respectivement associés à une faible et une forte activité épileptique, et de décrire le nombre de crises dans chaque état par des distributions poissonniennes.

En médecine, et dans le cadre des essais cliniques notamment, il est fréquent que les données recueillies proviennent de différents sujets. Tout en imaginant pouvoir expliquer chaque jeu de données individuel au moyen d'une structure de Markov sous-jacente, la complexité des données requiert néanmoins une démarche d'analyse particulière. En effet, la manifestation des symptômes fluctue d'un individu à l'autre. L'interprétation rigoureuse de cette variabilité entre les sujets nécessite d'adopter une approche populationnelle, pour laquelle l'utilisation de modèles mixtes est requise. Ce problème, discuté par Altman (2005), a récemment motivé la définition des modèles de Markov cachés à effets mixtes (*Mixed-effects hidden Markov models, MHMM*).

La définition des modèles de Markov cachés à effets mixtes, due à Altman (2007), s'opère en deux temps. D'abord, autant de modèles de Markov cachés qu'il y a de jeux de données individuels $\mathbf{y}_i$ sont établis. On notera $\mathbf{z}_i$ la séquence d'états cachés pour l'individu

i, et $\mathbf{\Psi}_i$ le jeu de paramètres intervenant dans la spécification du i^{eme} *HMM* individuel (probabilités de transition, probabilités d'émission). Dans un second temps, les $\mathbf{\Psi}_i$ sont définis comme des variables aléatoires:

$$\mathbf{\Psi}_i = h\left(C_i, \mu, \eta_i\right)$$

où

$$\eta_i \underset{i.i.d.}{\backsim} \mathcal{N}(0, \Omega)$$

et C_i est une matrice de covariables pour le sujet i. μ désigne l'ensemble des effets fixes du modèle. Ils permettent de décrire l'évolution moyenne de la maladie dans la population étudiée. La matrice de variance-covariance Ω permet de rendre compte de la variabilité des manifestations symptômatiques entre les sujets. L'ensemble des paramètres μ et Ω sont qualifiés de paramètres de population. On notera $\theta = (\mu, \Omega)$.

L'apprentissage des *MHMM* n'est pas direct. Par la structure-même du modèle, ce problème s'apparente à un problème de données incomplètes, pour lequel la vraisemblance s'exprime difficilement. Nous proposons la démarche d'apprentissage suivante. Pour commencer, l'algorithme MCMC-SAEM sera adapté aux modèles de Markov cachés à effets mixtes pour en estimer les paramètres de population du maximum de vraisemblance:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ p\left(\mathbf{y}_1, \ldots, \mathbf{y}_n; \theta\right)$$

Chacune des itérations de l'algorithme mettra à profit la procédure forward, qui propose une méthode de calcul rapide de la vraisemblance dans les *HMM*. Ensuite, cette première estimation à l'échelle populationnelle nous permettra d'établir les paramètres propres à chaque sujet par maximisation a posteriori de leur distribution:

$$\hat{\mathbf{\Psi}}_i = \underset{\mathbf{\Psi}_i}{\operatorname{argmax}}\ p\left(\mathbf{\Psi}_i | \mathbf{y}_i; \hat{\theta}\right)$$

Enfin, la simple application de l'algorithme de Viterbi dans les modèles de Markov cachés individuels nous donnera les séquences d'états les plus probables pour chaque individu:

$$\hat{\mathbf{z}}_i = \underset{\mathbf{z}_i}{\operatorname{argmax}}\ p\left(\mathbf{z}_i | \mathbf{y}_i; \hat{\mathbf{\Psi}}_i\right)$$

Des études de Monte-Carlo apportent des résultats encourageants quant aux performances des estimateurs obtenus. Enfin, cette nouvelle méthodologie a été mise en œuvre sur la problématique de l'épilepsie.

Bibliographie

[1] Albert (1991), A two state Markov mixture model for a time series of epileptic seizure counts, Biometrics.

[2] Albert and al. (1994), Time series for modelling counts from relapsing remitting disease: application to modelling disease activity in multiple sclerosis, Statistics in Medicine.
[3] Altman, Petkau (2005), Application of hidden Markov models to multiple sclerosis lesion count, Statistics in Medicine.
[4] Altman (2007), Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting, Journal of the American Statistical Association.
[5] Anisimov (2007), Analysis of responses in migraine modelling using hidden Markov models, Statistics in Medicine.
[6] Ip and al. (2007), Mixed effects hidden Markov models, Statistics in Medicine.
[7] Kuhn and Lavielle (2004), Coupling a Stochastic Approximation Version of EM with an MCMC Procedure, ESAIM: Probability and Statistics.
[8] Kuhn and Lavielle (2005), Maximum likelihood estimation in nonlinear mixed effects models, Computational Statistics and Data Analysis.
[9] Rabiner (1989), A tutorial on Hidden Markov Models and selected applications in speech recognition,Proceedings of the IEEE.