Modèle de fragilité et algorithme EM stochastique

Charles El-Nouty, Estelle Kuhn

Modèle de fragilité et algorithme EM stochastique.

Charles El-Nouty et Estelle Kuhn

UMR 557 Inserm/Inra/Cnam/Paris 13, SMBH - Université Paris 13, 74, rue Marcel Cachin, 93017 Bobigny, France ; c.el-nouty@uren.smbh.univ-paris13.fr & MIA, INRA, Domaine de Vilvert, 78352 Jouy-en-Josas, France ; estelle.kuhn@jouy.inra.fr.

Résumé en français

Le modèle de Cox (1972) joue un rôle essentiel en analyse de survie en particulier dans le cadre des études épidémiologiques. Une hypothèse classique sous-jacente est l'indépendance des données, au moins conditionnellement aux covariables. Mais cette hypothèse n'est souvent pas vérifiée du fait de l'hétérogénéité intrinsèque à la population observée, par exemple si les observations peuvent être regroupées par familles ou par zones géographiques. Le modèle de fragilité, qui prend en compte cette hétérogénéité en considérant un effet aléatoire, a été introduit dans l'article de Vaupel et al. (1979).

Il existe une vaste littérature sur l'estimation des paramètres dans le modèle de fragilité (cf. les ouvrages de Hougaard (2000) et Duchateau et Janssen (2008)). En général, on considère la vraisemblance observée et l'estimateur du maximum de vraisemblance associé. Les propriétés asymptotiques de cet estimateur ont d'ailleurs été établies. Mais il est souvent difficile voir impossible de le calculer directement. Pour résoudre ce problème, il existe principalement deux types d'approches. La première consiste à appliquer au modèle de fragilité les idées de Cox en recherchant une vraisemblance approchée. La seconde consiste en une approximation numérique de l'estimateur au lieu d'un calcul direct. Les variables aléatoires de fragilité étant non observées, on peut se placer dans le contexte plus général des modèles à variables latentes. Pour approcher l'estimateur du maximum de vraisemblance observée dans de tels modèles, on a généralement recours à l'algorithme Expectation Maximization (EM) proposé dans l'article de Dempster et al. (1977). Les propriétés de convergence de cet algorithme ont été prouvées. Cependant il n'est pas applicable dans beaucoup de cas pratiques en particulier à cause de la première étape qui requiert un calcul d'espérance conditionnelle souvent difficile à mener analytiquement. Plusieurs auteurs ont donc proposé des approximations de cet algorithme. Beaucoup d'entre elles, certes plus faciles à mettre en oeuvre, ne conservent cependant pas les propriétés de convergence de l'algorithme EM initial. D'autres font appel à des méthodes stochastiques qui demandent d'importants temps de simulation.

Dans El-Nouty et Kuhn (2010), nous proposons un algorithme convergent pour l'estimation par maximum de vraisemblance dans le modèle de fragilité. Nous considérons une approximation stochastique de l'algorithme EM couplé à une méthode de Monte Carlo par chaînes de Markov (SAEM-MCMC) proposée dans l'article de Kuhn et Lavielle (2004).

Dans cet algorithme, l'étape de calcul d'espérance conditionnelle est remplacée par deux étapes consistant d'abord en la simulation des variables de fragilité non observées qui sont ensuite utilisées dans une approximation stochastique de la log-vraisemblance complète. L'étape de maximisation reste la même. La suite générée par l'algorithme SAEM-MCMC converge presque sûrement vers un maximum local de la vraisemblance sous des hypothèses peu contraignantes. De plus, cet algorithme reste rapide, bien que faisant appel à la simulation, du fait de l'imbrication astucieuse des deux méthodes SAEM et MCMC.

Nous considérons ici une population d'individus regroupés en N groupes. Pour $1 \leq i \leq N$, on désigne par n_i la taille du i-ième groupe. Pour $1 \leq i \leq N$ et $1 \leq j \leq n_i$ le temps d'intérêt et le temps de censure de l'individu j du groupe i sont modélisés par des variables aléatoires notées T_{ij} et C_{ij} respectivement. On définit les variables aléatoires $Y_{ij} = \min(T_{ij}, C_{ij})$ et $\Delta_{ij} = \mathbb{1}_{\{T_{ij} \leq C_{ij}\}}$, où $\mathbb{1}_A$ désigne la fonction indicatrice de l'ensemble A. On observe alors les couples (Y_{ij}, Δ_{ij}). Pour $1 \leq i \leq N$, on note b_i le vecteur aléatoire de fragilité de dimension q pour le groupe i. On suppose que les vecteurs aléatoires (b_i) sont indépendants et identiquement distribués. Soient β un vecteur inconnu de dimension p et λ_0 une fonction de hasard inconnue. Dans la suite, nous faisons l'hypothèse que la fonction λ_0 est paramétrique ; on note α le paramètre vectoriel associé. Soient x_{ij} (respectivement z_{ij}) un vecteur de design de dimension p (respectivement q). On note $\lambda_{ij}(t|b_i)$ le risque conditionnel instantané du j-ième individu du i-ième groupe au temps t. Le modèle de fragilité s'écrit alors :

$$\lambda_{ij}(t|b_i) = \lambda_0(t) \exp(x_{ij}^t \beta + z_{ij}^t b_i), \quad 1 \leq i \leq N,\ 1 \leq j \leq n_i,\ t \geq 0, \tag{1}$$

où ^{t} désigne l'opérateur de transposition.

On note $\mathbf{y}, \boldsymbol{\delta}$ et $\mathbf{b}$ les vecteurs d'observations correspondant respectivement à $(y_{ij}), (\delta_{ij})$ et (b_i). On désigne par f_η la densité de probabilité des variables de fragilité. La vraisemblance complète L_N s'écrit alors :

$$L_N(\mathbf{y}, \boldsymbol{\delta}, \mathbf{b}; \theta) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \left(\lambda_{ij}(y_{ij}|b_i)^{\delta_{ij}} \exp\left(- \int_0^{y_{ij}} \lambda_{ij}(u|b_i) du \right) f_\eta(b_i) \right). \tag{2}$$

On considère l'estimateur du maximum de vraisemblance pour les paramètres $\theta = (\beta, \alpha, \eta)$, c'est-à-dire la valeur $\hat{\theta}_N$ de θ qui maximise la vraisemblance observée définie par :

$$L_N^{obs}(\mathbf{y}, \boldsymbol{\delta}; \theta) = \int L_N(\mathbf{y}, \boldsymbol{\delta}, \mathbf{b}; \theta) d\mathbf{b}. \tag{3}$$

Pour approcher $\hat{\theta}_N$, on utilise l'algorithme SAEM-MCMC proposé par Kuhn et Lavielle (2004). La k-ième itération consiste en trois étapes :

Etape 1 : Les variables de fragilité non observées $\mathbf{b}$ sont simulées selon une probabilité de transition Π_θ d'une chaîne de Markov convergente ayant comme distribution stationnaire la loi a posteriori $\pi_\theta(.|\mathbf{Y}, \mathbf{\Delta})$:

$$\mathbf{b}^k \sim \Pi_{\theta_{k-1}}(\mathbf{b}^{k-1}, \cdot)\,.$$

Etape 2 : On effectue une approximation stochastique de la log-vraisemblance complète en utilisant les valeurs simulées :

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_{k-1}\,\Big(\log L_N(\mathbf{Y}, \mathbf{\Delta}, \mathbf{b}^k; \theta) - Q_{k-1}(\theta)\Big),$$

où $\boldsymbol{\gamma} = (\gamma_k)_k$ est une suite de pas positifs décroissante.

Etape 3 : Les paramètres sont mis à jour dans l'étape de maximisation :

$$\theta_k = \arg\max Q_k(\theta)\,.$$

Les valeurs initiales Q_0 et θ_0 sont choisies arbitrairement. La suite de pas $(\gamma_k)_k$ est habituellement choisie de sorte que $\gamma_k = \frac{1}{k^\mu}$ avec $\frac{1}{2} < \mu \leq 1$. Les probabilités de transition Π_θ peuvent être obtenues par un algorithme de Metropolis-Hastings ou par un échantillonneur de Gibbs. La suite (θ_k) générée par l'algorithme SAEM-MCMC converge presque sûrement vers un maximum local de la vraisemblance observée sous des hypothèses peu contraignantes (cf. Kuhn et Lavielle (2004)).

Pour les études numériques simulées, nous considérons le modèle suivant :

$$\lambda_{ij}(t|b_i) = \lambda_0(t)\exp\Big(b_{0i} + x_{ij}^t(\beta + b_{1i})\Big) \quad 1 \leq i \leq N,\ 1 \leq j \leq n_i,\ t \geq 0, \tag{4}$$

où les variables (b_{0i}) (respectivement (b_{1i})) sont indépendantes et identiquement distribuées de loi normale $\mathcal{N}(0, \sigma_0^2)$ (respectivement $\mathcal{N}(0, \sigma_1^2)$) et (b_{0i}) et (b_{1i}) sont indépendantes. Les covariables (x_{ij}) sont binaires. Les suites (θ_k) obtenues pour l'estimation des trois paramètres $(\beta, \sigma_0^2, \sigma_1^2)$ lors d'une trajectoire de l'algorithme sont présentées en Figure 1.

Nous comparons nos résultats numériques à ceux obtenus par d'autres algorithmes développés dans la littérature. Nous projetons également d'appliquer cette méthode sur des données réelles.

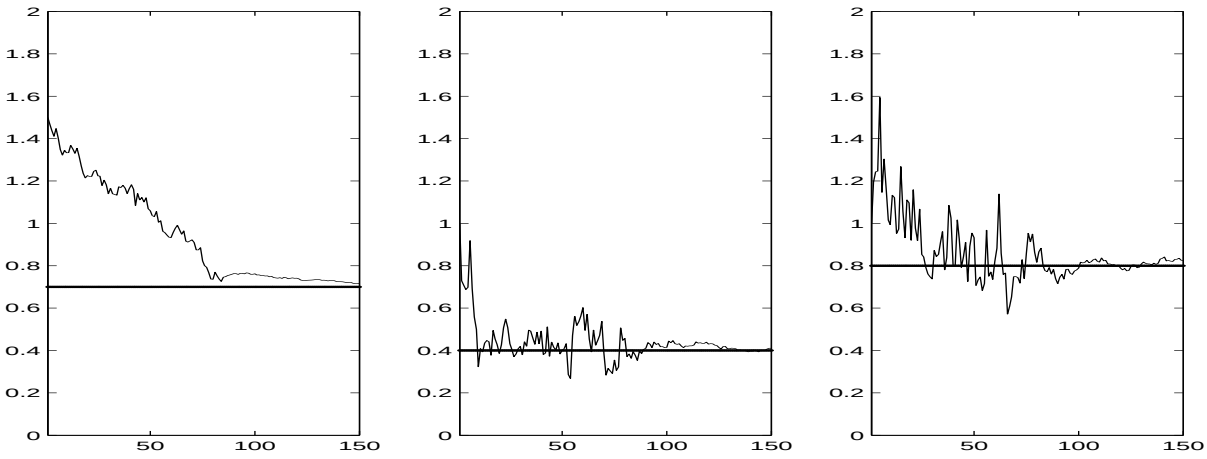

Figure 1: Estimation des paramètres $(\beta, \sigma_0^2, \sigma_1^2)$: présentation d'une trajectoire de l'algorithme SAEM-MCMC. A gauche : estimation de β, au centre : estimation de σ_0^2, à droite : estimation de σ_1^2 (vraies valeurs des paramètres $(0.7, 0.4, 0.8)$; initialisations $(2, 1, 1)$).

Résumé en anglais

The Cox (1972) proportional hazards model plays a key role in survival analysis and particularly in many epidemiological problems. A classical underlying assumption is that the observations are independent, at least conditionally upon covariates. But this assumption is often not fulfilled because of the lack of homogeneity for the population of interest. To measure this unobserved heterogeneity, frailty models were introduced by Vaupel and al. (1979).

There is a huge literature on estimation procedures for the parameters for the frailty models. The basic idea consists in considering the observed likelihood and to solve the corresponding likelihood equations. Indeed asymptotic theoretical convergences for the Maximum Likelihood Estimator (MLE) were established. But it is often not possible in many practical cases to compute directly the MLE. To overcome this crucial difficulty, two main approaches are developed. The first one consists in applying Cox's idea to the frailty models for obtaining an approximated observed likelihood. The second one consists in a numerical approximation of the MLE instead of a direct computation. Since the frailties are not observed, the underlying model belongs to the family of models with hidden variables. Dempster and al. (1977) proposed the used of the Expectation Maximization (EM) algorithm to solve the MLE problem in such models. The theoretical convergence of this algorithm was established. However in many cases the EM algorithm can not be directly applied particularly the expectation step. Thus several authors suggest approximations of this algorithm. Some of them were very time consuming since they use

intensive simulation processes. Some others are more efficient in computation time but they have lost the theoretical convergence property of the initial EM algorithm.

In El Nouty and Kuhn (2010) we propose a new numerical convergent algorithm for maximum likelihood estimation in frailty models. We consider the Stochastic Approximation EM with Monte Carlo Markov Chain (SAEM-MCMC) algorithm introduced by Kuhn and Lavielle (2004). The flavor of this method is that the initial expectation step of the EM algorithm is divided into two new steps: the first one consists in simulating the non observed frailties whereas the second one computes a stochastic approximation of the complete log-likelihood by using the simulated values of the frailties. The maximization step follows the same lines as those of the EM algorithm. The SAEM-MCMC algorithm converges as surely towards the MLE under weak regularity conditions. Moreover the convergence is very quick, decreasing the computation time.

Mots clés

Données de survie et données censurées - Médecine - épidémiologie

Bibliographie

[1] Cox, D.R. (1972) Regression models and life-tables, J.R.S.S. Serie B, 34, 187-220.
[2] Vaupel, J.W., Manton, K.G. et Stallard, E. (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality, Demography, 16, 439-454.
[3] Hougaard, P. (2000) Analysis of multivariate survival data, Statistics for Biology and Health, Springer-Verlag, New York.
[4] Duchateau, L. et Janssen, P. (2008) The Frailty Model, Statistics for Biology and Health, Springer-Verlag, New York.
[5] Dempster, A.P., Laird, N.M. et Rubin, D.B. (1977) Maximum likelihood from incomplete date via the EM algorithm, J.R.S.S. Serie B, 39(1), 1-38.
[6] El-Nouty, C. and Kuhn, E. (in preparation) Maximum likelihood inference in frailty model via a convergent stochastic EM algorithm.
[7] Kuhn, E. et Lavielle, M. (2004) Coupling a stochastic approximation version of EM algorithm with an MCMC procedure, ESAIM P&S, 8,115-131.