Analyse bayésienne de données de survie discrètes sujettes à censure informative avec un modèle à effets aléatoires partagés reparamétré.

Sophie Ancelet-Enjalric, Elise de la Rochebrochard, Jean Bouyer

Analyse bayésienne de données de survie discrètes sujettes à censure informative avec un modèle à effets aléatoires partagés reparamétré.

Sophie Ancelet (a)(b) & Elise de la Rochebrochard (b)(c)(d) & Jean Bouyer (b)(c)(d)

(a) INRA, Département MIA, Unité Mét@risk, AgroParistech, Paris, France
(b) CESP-INSERM U1018, Epidémiologie de la reproduction et du développement de l'enfant, Hôpital du Kremlin-Bicêtre, France
(c) Université Paris Sud 11, UMRS 1018, F-94276 Le Kremlin-Bicêtre, France
(d) INED, F-75020, Paris, France

Abstract

Les mécanismes de censure dits informatifs viennent souvent complexifier l'analyse de données de survie. L'inférence de modèles standards ne tenant pas compte de ce type de données manquantes peut mener à des conclusions biaisées. De nombreux travaux se sont axés autour des modèles à effets aléatoires partagés pour l'analyse de données longitudinales avec censure informative. Nous présentons une extension à cette classe de modèles hiérarchiques pour l'analyse de données de survie discrètes sujettes à censure informative. Notre modèle est basé sur la combinaison de deux modèles à hasards proportionnels en temps discret et tient compte de l'existence possible d'une variabilité résiduelle non-partagée. Puis, nous proposons d'utiliser une reparamétrisation du modèle proposé, par ajout de paramètres redondants, ainsi qu'un choix spécifique de lois a priori afin d'améliorer les propriétés de convergence des algorithmes MCMC implémentés pour mener l'inférence bayésienne du modèle. Enfin, nous proposons de valider notre modèle à partir de calculs de facteurs de Bayes partiels en vue de tester l'hypothèse d'existence d'une censure informative intervenant sur un mécanisme de survie d'intérêt et le calcul de p-valeurs bayésiennes "mixtes" en vue de quantifier l'adéquation entre notre modèle et des données observées. Nous illustrons la pertinence de notre approche bayésienne par des simulations ainsi que sur un jeu de données réelles issues de l'enquête "Devenir Après Initiation de la Fécondation-In-Vitro".

Key-Words: censure informative, données de survie discrètes, facteur de Bayes partiel, inférence bayésienne, modèle à effets aléatoires partagés, p-valeur bayésienne "mixte", reparamétrisation par ajout de paramètres

Abstract

Survival data analysis may often be complicated by the existence of one (or several) informative censoring mechanism(s). Estimation procedures without adjusting for this type of missing data may lead to biased conclusions. Many studies have focused on shared random effect models to analyse longitudinal data subject to informative censoring. We present an extension of this class of hierarchical models to analyze discrete survival data subject to informative censoring. Our model is based on the combination of two discrete proportional hazards models and takes into account that a non-shared residual variability may exist. Then, we propose to use a parameter expansion and to assign specific prior distributions on the variance parameters to improve the convergence properties of the MCMC algorithms implemented to infer our model. Finally, we propose to validate our model by computing partial Bayes factors in order to test if an informative censoring has an influence on a survival mechanism of interest and "mixte" Bayesian p-values in order to quantify the adequation between our model and observed data. We illustrate the efficiency of our Bayesian approach from simulation studies and from real data coming from the scientific survey "Devenir Après Initiation de la Fécondation-In-Vitro".

Key-Words: informative missingness, discrete survival data, partial Bayes factor, Bayesian inference, shared random effects model, "mixte" Bayesian p-value, parameter expansion

1 Introduction

De nombreuses études de cohortes s'intéressent au délai d'occurrence d'un événement aléatoire particulier qui, lorsqu'il se produit, n'apparait qu'une seule fois au cours de la période de suivi des sujets: guérison, 1er accouchement... Des mécanismes de censure dits informatifs viennent souvent complexifier l'analyse de telles données de survie. On parle de censure informative lorsque la probabilité d'occurrence de données manquantes dûes à un tel mécanisme de censure dépend de l'événement principal d'intérêt [1]. Dans ce contexte, une non prise en compte de l'information potentielle apportée par l'occurrence (ou non) de données manquantes peut notamment conduire à une estimation biaisée de la probabilité *théorique* (i.e., dans une population sans observation manquante) d'occurrence de l'événement d'intérêt.

Récemment, de nombreux travaux se sont axés autour des modèles à effets aléatoires partagés pour (i) l'analyse de données longitudinales avec censure informative [2] (ii) l'analyse simultanée de données longitudinales et de données de survie [3]. Dans un premier temps, nous proposons une extension de la classe des modèles à effets aléatoires partagés pour l'analyse de données de survie discrètes sujettes à censure informative. A notre connaissance, l'utilisation de cette classe de modèles n'a pas été envisagée dans ce contexte particulier malgré sa justification souvent très intuitive. Dans le domaine biomédical, par exemple, une hypothèse de modélisation plausible est que l'événement

aléatoire d'intérêt (e.g., guérison, 1er accouchement) et le mécanisme aléatoire de censure (e.g., abandon en cours de traitement, décès) dépendent simultanément d'un même effet aléatoire latent: l'état de santé du patient.

L'estimation et la validation bayésienne des modèles à effets aléatoires partagés ont été très peu abordées dans la littérature pour l'analyse de données longitudinales avec censure informative ou encore l'analyse jointe de données longitudinales et données de survie. Pourtant, pour de tels modèles dont l'estimation nécessite de nombreuses intégrations multi-dimensionnelles, l'approche bayésienne présente souvent de nombreux avantages [4] par rapport aux algorithmes EM implémentés dans le cadre classique. Dans ce travail, nous proposons d'utiliser une reparamétrisation du modèle à effets aléatoires partagés proposé, par ajout de paramètres multiplicatifs redondants, ainsi qu'un choix spécifique de lois a priori non-informatives permettant une estimation bayésienne plus efficace du modèle (meilleure visite de l'espace des paramètres, estimations plus fiables). En outre, nous proposons de valider notre modèle à partir du (i) calcul de facteurs de Bayes partiels permettant de tester l'hypothèse d'existence d'une censure informative intervenant sur le mécanisme de survie d'intérêt (ii) calcul de p-valeurs bayésiennes "mixtes" [5] basées sur la fonction de discrépance du ξ^2 permettant d'évaluer la capacité de notre modèle à reproduire des données de survie censurées plausibles au vue des données observées.

Nous illustrons la pertinence de notre approche bayésienne sur des simulations ainsi que les données de survie censurées discrètes issues de l'enquête "Devenir Après Initiation de la Fécondation-In-Vitro" (DAIFI).

2 Modèle et notations

Dans ce travail, nous considérons le cas de données de survie censurées à droite par une décision d'abandon des sujets pendant la période de suivi (pas de données manquantes intermittentes). Soit une cohorte de I sujets suivis pendant, au plus, J temps d'observation discrets (J fixé supposé petit). Pour chaque sujet i (i=1,2,. . .,I), nous nous intéressons au premier temps d'occurrence T_i d'un événement aléatoire d'intérêt et au temps aléatoire Z_i à partir duquel le sujet est définitivement exclu de la cohorte suite à une décision d'abandon. $T_i = J + 1$ et $Z_i = J$ lorsque, respectivement, l'événement d'intérêt ne s'est pas produit et le sujet n'a pas abandonné sur la période de suivi considérée.

Pour tout i=1,. . .,I, nous définissons les probabilités conditionnelles suivantes:

$$p_{ij} = Pr(T_i = j | T_i > j-1; f_i) \qquad \forall j = 1, \ldots, J$$
$$\pi_{il} = Pr(Z_i = l | Z_i > l-1; f_i, \epsilon_i) \qquad \forall l = 1, \ldots, J-1$$

p_{ij} est la probabilité conditionnelle d'occurrence de l'événement d'intérêt, au temps j, pour le sujet i sachant qu'il ne s'est pas produit précédemment. π_{il} est la probabilité conditionnelle que le sujet i abandonne après l'observation l sachant qu'il n'a pas abandonné précédemment.

Nous proposons de décomposer les probabilités conditionnelles ci-dessus, à échelle logit, selon la formulation assymétrique suivante:

$$logit(p_{ij}) = \beta^T X_{ij} + f_i$$
$$logit(\pi_{il}) = \rho^T W_{il} + \lambda f_i + \epsilon_i$$

où β^T et ρ^T sont des vecteurs de log odds-ratios correspondant aux covariables X_{ij} et W_{il} respectivement; les f_i sont des effets aléatoires jouant le rôle de substituts pour tous les facteurs déterminants non-observés et spécifiques à chaque sujet i qui expliquent la probabilité conditionnelle d'occurrence de l'événement d'intérêt et influent simultanément sur la probabilité conditionnelle d'abandon; les ϵ_i sont des effets aléatoires résiduels, propres au comportement d'abandon des sujets. Ils jouent le rôle de substituts pour tous les facteurs déterminants non-observés susceptibles d'expliquer une variabilité résiduelle de la probabilité conditionnelle d'abandon, non-partagée avec la probabilité conditionnelle d'occurrence de l'événement d'intérêt. Enfin, λ est un paramètre réel jouant le rôle de coefficient de regression. Il permet de quantifier l'effet de la variable latente f_i sur la probabilité conditionnelle d'abandonner (à échelle logit). Ce paramètre a un intérêt majeur dans notre modélisation. En effet, une valeur non-nulle de λ induit une dépendance entre l'occurrence de l'événement d'intérêt et le comportement d'abandon: cela correspond à l'existence d'une censure informative. Au contraire, si $\lambda = 0$ alors cela signifie qu'il n'existe pas de censure informative induite par des facteurs spécifiques aux sujets de la cohorte. Afin de modéliser l'hétérogénéité inter-sujets, nous assignons les lois a priori échangeables suivantes sur les effets aléatoires du modèle: $f_i \sim^{i.i.d} N(0, \sigma_f^2)$ et $\epsilon_i \sim^{i.i.d} N(0, \sigma_\epsilon^2)$.

3 Inférence bayésienne: expansion de paramètres et choix des lois a priori

Dans le cadre d'une inférence bayésienne, nous cherchons à évaluer la distribution a posteriori de tous les paramètres et variables latentes du modèle proposé. Ces distributions a posteriori n'ayant pas une forme analytique connue, nous avons recours à un algorithme MCMC pour étudier la loi a posteriori des paramètres. L'algorithme MCMC à implémenter ne pose pas de difficulté particulière. En revanche, la convergence de l'algorithme s'avère souvent difficile, en particulier pour les paramètres de variance pour lesquels un fort niveau d'autocorrélation intra-chaînes peut se manifester indiquant une visite lente de l'espace des paramètres. Cela peut s'expliquer par la structure relativement contrainte des modèles à effets aléatoires partagés à laquelle peut venir s'ajouter le problème bien connu de l'estimation bayésienne par algorithmes MCMC de paramètres de variance proches de zéro dans le cas de modèles hiérarchiques [6].

Dans ce contexte, nous suggérons d'étendre au modèle à effets aléatoires partagés précédemment proposé, l'approche par expansion de paramètres proposée par Kinney et

Dunson (2007) [7]. Basée sur une décomposition de Cholesky modifiée de la matrice de variance-covariance des effets aléatoires, cette approche revient à écrire:

$$logit(p_{ij}) = \beta^T X_{ij} + \alpha\xi_i$$
$$logit(\pi_{il}) = \rho^T W_{il} + \lambda\alpha\xi_i + \gamma\omega_i$$

où α et ω sont deux paramètres multiplicatifs redondants. Bien que non-identifiables au vu des données observées, ces nouveaux paramètres doivent permettre de réduire les dépendances potentielles entre paramètres à inférer sans détériorer leur estimation.

Nous proposons d'utiliser des lois a priori standards i.e., normales plates centrées pour les effets fixes. Suivant les idées de Kinney et Dunson (2007), nous suggérons par ailleurs d'utiliser les lois a priori suivantes $\alpha \sim N(0,1)$, $\gamma \sim N(0,1)$, $\xi_i \sim^{i.i.d} N(0,\sigma_\xi^2)$ et $\omega_i \sim^{i.i.d} N(0,\sigma_\omega^2)$. Cela revient à assigner une loi a priori demi-Cauchy propre, reconnue pour être souvent plus robuste par rapport à la famille conjuguée des lois Inverse-Gamma [8], sur les écarts-types σ_ϵ et σ_f. Le choix d'une loi a priori sur le paramètre λ est plutôt délicat. Aussi, nous avons testé des lois a priori propres de formes différentes (e.g., Unif(-a,a), Normal(0,σ^2_{prior})) afin de tester la robustesse des résultats obtenus par rapport au choix de cette loi a priori.

Nous avons évalué par simulations l'impact de cette reparamétrisation par ajout de paramètres multiplicatifs redondants sur l'estimation bayésienne des paramètres λ, σ_f et σ_ϵ. Comme nous pouvions nous y attendre, les estimateurs ponctuels obtenus pour les paramètres λ, σ_f et σ_ϵ sont très similaires. En revanche, une réduction significative des corrélations intra-chaînes apparaît et par conséquent, une bien meilleure exploration de l'espace des paramètres par les chaînes de Markov.

4 Application

DAIFI est une enquête scientifique menée par l'INSERM (Institut National de la Santé et de la Recherche Médicale) et par l'INED (Institut national d'Etudes Démographiques) dont l'objet est l'étude du parcours des couples pendant et après un traitement par Fécondation-In-Vitro (FIV). Dans le cadre de cette enquête, nos objectifs sont (1) d'estimer la probabilité *théorique* (i.e., dans une population sans abandon) qu'un couple, débutant un programme de FIV, obtienne une naissance après au maximum quatre tentatives successives en tenant compte de l'information potentielle apportée par un comportement d'abandon (2) de valider l'hypothèse selon laquelle il existerait un lien de dépendance stochastique entre la probabilité d'occurrence d'une naissance par FIV et la probabilité d'abandon d'un couple en cours de traitement.

Nous considérons une cohorte de 3002 femmes ayant débuté un programme de FIV entre 1998 et 2002 dans deux centres de traitement différents: Cochin ou Clermont-Ferrand. Pour chaque femme, nous nous limitons à une période de suivi courte en temps discret:

4 tentatives i.e., les 4 ponctions d'ovocytes successives prises en charge par la sécurité sociale. Pour chaque femme i, nous nous intéressons au temps d'occurrence du premier accouchement sur les 4 tentatives considérées. Plus de 50% des femmes ont abandonné, i.e. interrompu leur traitement, avant la naissance souhaitée, sur les 4 tentatives considérées. Les experts du domaine évoquent l'hypothèse selon laquelle les femmes qui abandonnent en cours de traitement ont généralement de mauvais facteurs pronostics (c.a.d niveau de fertilité) impliquant, de fait, une faible probabilité d'accouchement après un traitement par FIV, d'où un recours intuitif aux modèles à effets aléatoires partagés.

Nous avons mené l'inférence bayésienne du modèle à effets aléatoires partagés proposé à partir des données de survie discrètes de l'enquête DAIFI. Dans un premier temps, nous avons tester l'hypothèse d'existence d'un lien de dépendance stochastique significatif, induit par les facteurs pronostics des femmes et possiblement lié au centre de traitement choisi, entre la probabilité d'occurrence d'un accouchement et la probabilité d'abandon au cours d'un programme de FIV. Sous le paradigme bayésien, cela nous a amené à un problème de choix de modèle. Un calcul de facteurs de Bayes partiels, basé sur le choix d'échantillons d'apprentissage et d'échantillons de validation, nous a permis de comparer plusieurs modèles en compétition à partir d'un niveau d'information a priori équivalent à la spécification d'un prior informatif. Nos résultats valident l'hypothèse d'existence d'une censure informative, quelque soit le centre de traitement considéré, liée aux facteurs pronostics des femmes. Par ailleurs, notre modèle met en évidence l'existence d'une corrélation négative beaucoup plus forte entre probabilité d'occurrence d'un accouchement et probabilité d'abandon pour le centre de Clermont-Ferrand par rapport au centre de Cochin (intervalles de crédibilité à 95%: λ_{CF} [-0.43,0.04], λ_{Cochin} [-0.20,0.24]). Ce résultat intéressant pourrait s'expliquer par la localisation géographique des centres choisis et plus précisément par le fait que le choix offert aux femmes en terme de centre de traitement par FIV est beaucoup plus large autour du centre de Cochin (Paris) qu'à Clermont-Ferrand. De ce fait, les femmes ayant de bons facteurs pronostics sont susceptibles d'abandonner en cours de programme pour se rendre dans un autre centre de traitement. Dans un deuxième temps, nous avons déduit de notre modèle un estimateur de la probabilité *théorique* qu'un couple, débutant un programme de FIV, obtienne une naissance après au maximum quatre tentatives successives. Nous mettons en évidence les éventuels biais d'estimation de cette probabilité induits par des approches ne tenant pas compte de l'existence d'une censure informative significative liée aux facteurs pronostics des femmes. Nous chercherons à quantifier l'adéquation de notre modèle à effets aléatoires partagés par rapport aux données observées, en utilisant des distributions prédictives a posteriori "mixtes" permettant le calcul de p-valeurs bayésiennes reconnues pour être moins conservatives que les p-valeurs bayésiennes classiques [5].

Bibliographie

[1] Wu, M.C. et Carroll, R.J. (1998) Estimation and comparisson of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 45,

175–188
[2] Schluchter, M.D. (1992) Methods for the analysis of informatively censored longitudinal data. *Statistics in medicine*, 11, 1861–1870
[3] Henderson, R., Diggle, P. Dobson, A. (2000) Joint modelling of longitudinal measurements and event time data . *Biostatistics*, 1(4), 465–480
[4] Hu, W., Li, G., Li, N. (2009) A Bayesian approach to joint analysis of longitudinal measurements and competing risks failure time data. *Statistics in medicine*, 28, 1601–1619
[5] Marshall, E. C., Spiegelhalter, D. J. (2003) Approximate cross-validatory predictive checks in disease mapping. *Statistics in medicine*, 22, 1649-1660
[6] Gelman, A. (2004) Parametrization and Bayesian Modeling *Journal of the American Statistical Association*, 99,537-545
[7] Kinney, S. K., Dunson, D. B. (2007) Fixed and Random Effects Selection in Linear and Logistic Models *Biometrics*, 63, 690-698
[8] Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models *Bayesian Analysis*, 3, 515-533