Modèle de Poisson mixte à classes latentes avec sur-représentation de zéros: application à l'identification de trajectoires hétérogènes d'intensité d'exposition vie entière

Emilie Lévêque [1] & Karen Leffondré[1] & Cécile Proust-Lima [1]

[1] *Université de Bordeaux, ISPED, INSERM, Bordeaux Population Health Research Center, Biostatistics Team, UMR 1219*
emilie.leveque@u-bordeaux.fr

Résumé. De nombreuses études épidémiologiques tentent d'identifier des profils de trajectoires d'exposition avec la perspective d'estimer leur association avec le risque de survenue d'un évènement de santé, comme le risque de développer un cancer. Le modèle linéaire mixte à classes latentes permet de tenir compte de l'hétérogénéité dans les trajectoires de mesures répétées en identifiant des sous groupes d'individus ayant une évolution temporelle moyenne spécifique. Bien qu'il constitue une piste intéressante, ce modèle reste assez peu utilisé dans le cadre des expositions prolongées environnementales. Une des raisons est que les expositions environnementales ont généralement des distributions très particulières avec une large proportion d'intensités faibles ou nulles que les modèles linéaires mixtes à classes latentes classiques ne peuvent pas appréhender. Notre objectif était donc de proposer et implémenter un modèle de Poisson mixte à classes latentes avec sur-représentation de zéros (ZIP-LCMM) pour gérer ce genre de données. L'inférence est réalisée par maximum de vraisemblance. L'intégrale sur les effets aléatoires présente dans la vraisemblance est approchée par une méthode de quadrature gaussienne pseudo-adaptative en deux étapes. Ce modèle est appliqué à une étude sur la consommation de cigarettes vie entière et le cancer du poumon. Ces données proviennent de l'étude cas-témoins française, ICARE, basée en population générale; seuls les hommes fumeurs ont été considérés ici (1938 cas / 1837 témoins). Le modèle a permis d'identifier des profils d'évolution de consommation de cigarettes bien distincts, tous associés à des risques de cancer du poumon différents. Avec le ZIP-LCMM, nous avons donc pu identifier des trajectoires d'exposition vie entière en prenant en compte une large proportion de zéros dans la distribution des mesures répétées. Ce développement méthodologique donne de nouvelles perspectives pour l'identification de trajectoires d'expositions environnementales vie entière.

Mots-clés. Données d'exposition, trajectoires vie entière, modèles à classes latentes, modèles mixtes, distribution de Poisson à sur-représentation de zéros, données longitudinales

Abstract. Many epidemiological studies attempt to identify longitudinal profiles of an exposure in order to estimate their association with the risk of an health outcome, such as the risk of having a cancer. The latent class linear mixed model takes into account heterogeneity in trajectories of repeated measures over time by identifying subgoups with a specific temporal mean evolution. Although it appears as a useful statistical tool, this model is barely used for protracted environmental exposures. One reason may be that environmental exposures usually have peculiar distributions with a large proportion of low or null intensities, that the latent class linear mixed model cannot apprehend. Our objective was thus to propose and implement a latent class Zero-Inflated Poisson mixed (ZIP-LCMM) model to handle such data. The model was estimated by maximum likelihood framework. The integral over the random effects present in the likelihood was approximated by a two-step pseudo-adaptive Gaussian quadrature rule. We applied this model to data on smoking exposure over lifetime and lung cancer. The data came from the ICARE French population-based case-control study; only male ever-smokers have been considered for the application (1938 cases / 1837 controls). The model allowed the identification of several profiles of smoking exposure associated with different risks of lung cancer. With ZIP-LCMM, we thus were able to identify lifetime exposure trajectories taking into account a large proportion of zeros in the distribution of repeated measures data. This methodological development gives new perspectives for the identification of environmental lifetime exposure trajectories.

Keywords. Exposure data, lifetime trajectories, latent class models, mixed models, Zero-Inflated Poisson distribution, longitudinal data

1 Contexte

De nombreuses études épidémiologiques tentent d'identifier des profils de trajectoires d'exposition avec la perspective d'estimer leur association avec le risque de survenue d'un évènement de santé, comme le risque de développer un cancer. Le modèle linéaire mixte à classes latentes permet de tenir compte de l'hétérogénéité dans les trajectoires de mesures répétées en identifiant des sous groupes d'individus ayant une évolution temporelle moyenne spécifique (Muthén (1999), Proust (2017)). Bien qu'il constitue une piste intéressante, ce modèle reste assez peu utilisé pour des expositions prolongées environnementales recueillies dans le cadre d'études cas-témoins. Pourtant, grâce au recueil de données rétrospectives auto-rapportées via des questionnaires standardisés, on peut disposer de mesures répétées quantitatives d'une exposition d'intérêt sur laquelle le sujet a été interrogé. Ainsi, il est possible d'obtenir des trajectoires individuelles d'une intensité d'exposition tout au long de l'histoire d'exposition du sujet.

Cependant, ces données d'expositions environnementales recueillies ont généralement des distributions très particulières avec une large proportion d'intensités faibles ou nulles au cours de l'histoire d'exposition des sujets (comprenant les périodes d'interruptions, les années depuis l'arrêt définitif, ...) que les modèles linéaires mixtes à classes latentes classiques ne peuvent pas appréhender. Pour pouvoir gérer convenablement cette proportion non négligeable de zéros dans un modèle à classes latentes, un modèle à sur-représentation de zéros (Lambert (1992)) a été considéré. De plus, considérer une distribution de Poisson pour les données répétées de telles expositions prolongées semble plus adaptée pour ce genre de données.

2 Objectif

L'objectif était de proposer et implémenter un modèle de Poisson à sur-représentation de zéros combiné à un modèle à classes latentes (ZIP-LCMM) pour identifier des trajectoires d'intensité d'exposition vie entière pour des données répétées comportant une large proportion de zéros. Une application de ce modèle a été réalisée pour identifier des profils de trajectoires d'intensité de consommation de tabac vie entière et d'en évaluer leur association avec le risque de cancer du poumon, à partir de données provenant d'une étude cas-témoins française sur les cancers respiratoires, nommée ICARE.

3 Méthode

3.1 Modèle ZIP-LCMM

Le modèle ZIP-LCMM considère un nombre fini G de sous-populations aux profils d'évolution différents. Il est composé de deux sous-modèles estimés simultanément. Le premier concerne la probabilité d'appartenance aux classes latentes tandis que le second caractérise la distribution individuelle des données répétées de la variable d'intérêt conditionnellement à chaque classe.

De manière similaire à un modèle linéaire mixte à classes latentes, l'appartenance à une classe latente est définie par une variable latente discrète c_i pour l'individu i, qui vaut g s'il appartient à la classe latente g ($g = 1, \ldots, G$). La probabilité de l'individu i d'appartenir à la classe latente g, notée π_{ig}, est obtenue à partir d'un modèle logistique multinomial (sous-modèle 1) :

$$\pi_{ig} = P(c_i = g|X_{pi}) = \frac{\exp(\zeta_{0g} + X_{pi}^T \zeta_{1g})}{\sum_{l=1}^{G} \exp(\zeta_{0l} + X_{pi}^T \zeta_{1l})} \quad (1)$$

où ζ_{0l} sont les intercepts spécifiques pour chaque classe l ($l \in \{1, \ldots, G\}$); ζ_{1l} sont les vecteurs des coefficients spécifiques à chaque classe l associés au vecteur de covariables X_{pi}^T. Par soucis d'identifiabilité du modèle, on fixe ζ_{0G}=0 et ζ_{1G}=0.

Le second sous-modèle (sous-modèle 2) décrit les profils moyens de trajectoires d'une variable Y à travers un modèle ZIP mixte spécifique à chaque classe latente. Conditionnellement à la classe latente g, les mesures répétées Y_{ij} d'un même individu i aux temps d'observation t_{ij} ($j = 1, \ldots, n_i$) suivent une distribution ZIP (Lambert (1992)). Une telle distribution est associée à deux processus (sous-modèle 2a et 2b) qui sont estimés en même temps. Le premier concerne la probabilité p_{ijg} d'être un zéro dit "structurel". Le second processus concerne les données répétées, qui sont représentées par une distribution de Poisson associée à son paramètre λ_{ijg}.

Le modèle s'écrit donc:

$$Y_{ij}|_{c_i=g} \sim ZIP(p_{ijg}, \lambda_{ijg})$$

où

- $p_{ijg} = P(\alpha_{ij} = 1|c_i = g)$ représente la probabilité spécifique à la classe g d'être un zéro structurel au temps j. Elle est représentée par α_{ij} qui est l'indicateur binaire qui vaut 1 si Y_{ij} est un zéro structurel, 0 sinon (Muthén (2005)).
  Ceci est modélisé par le modèle logit suivant (sous-modèle 2a) :

$$p_{ijg} = P(\alpha_{ij} = 1|c_i = g) = \frac{\exp(\varrho_{0g} + M_{ij}^T \varrho_{lg})}{1 + \exp(\varrho_{0g} + M_{ij}^T \varrho_{lg})} \tag{2}$$

  où ϱ_{0g} sont les intercepts spécifiques à chaque classe; M_{ij}^T est le vecteur des covariables associé au vecteur des coefficients spécifiques à chaque classe ζ_{lg}.

- $\lambda_{ijg} = E(Y_{ij}|_{c_i=g})$ est définie par un modèle mixte de Poisson (sous-modèle 2b) :

$$\ln(\lambda_{ijg}) = \beta_{0g} + X_{ij}^T \beta_{lg} + Z_{ij}^T b_i \tag{3}$$

  où X_{ij}^T est le vecteur des covariables associé au vecteur des coefficients des effets fixes β_{lg} spécifiques à chaque classe ; Z_{ij}^T est le vecteur des covariables associé au vecteur des effets aléatoires b_i.
  On suppose que $b_i|_{c_i=g} \sim \mathcal{N}(0, \sigma_{0g}^2 B)$ avec σ_{0g}^2 les coefficients de proportionnalité associés à la matrice de variance-covariance B. Par soucis d'identifiabilité du modèle, on fixe $\sigma_{0G}^2 = 1$.

3.2 Estimation par maximum de vraisemblance

Un tel modèle est estimé par maximum de vraisemblance pour un nombre de classes latentes G fixé en utilisant l'algorithme itératif de Marquardt (Proust (2017)). Le calcul de la log vraisemblance requiert le calcul d'une intégrale sur les effets aléatoires. Il n'y a pas de solution analytique à la vraisemblance individuelle. Ainsi, l'intégrale sur les effets aléatoires est approchée par une méthode de quadrature gaussienne pseudo-adaptative en deux étapes, s'appuyant sur les travaux de Rizopoulos (2012) et de Ferrer (2016). La stratégie d'estimation a dû être adaptée au contexte du modèle à classes latentes puisque cette méthode d'approximation a été validée et utilisée pour un modèle conjoint à effet aléatoire partagé (Rizopoulos (2012)) et un modèle conjoint multi-états (Ferrer (2016)).

Comme pour tout modèle à classes latentes estimé par maximum de vraisemblance, l'estimation du modèle doit être réalisée à partir de plusieurs jeux de valeurs initiales pour s'assurer de la convergence vers le maximum global. Le nombre optimal de classes latentes a été choisi en considérant à la fois un critère statistique (minimisation du BIC) mais également la pertinence des trajectoires estimées ainsi que l'effectif des sujets classés a posteriori composant chaque classe latente.
Le modèle ZIP-LCMM a été implémenté sous le logiciel R.

4 Simulation

Le programme d'estimation a été validé par une étude de simulations. En particulier, nous avons observé l'impact sur les estimations du modèle de la méthode d'intégration numérique en deux étapes proposée pour le calcul de vraisemblance.

5 Application

5.1 Données utilisées

Le modèle ZIP-LCMM a été utilisé pour identifier des profils de trajectoire vie entière de consommation de tabac et comparer le risque de cancer du poumon associé à chaque profil. Nous nous sommes basés pour cela sur des données réelles provenant de l'étude cas-témoins française sur les cancers respiratoires, ICARE, basée en population générale et conduite entre 2001 et 2007 (Luce (2006)). Dans le cadre de cette illustration, nous nous sommes seulement intéressés aux cas de cancer de poumon et aux hommes fumeurs (courant ou ancien) au moment de la date index - date diagnostic pour les cas et d'interview pour les témoins (sujet indemne de cancer du poumon au moment du diagnostic des cas). A partir des données recueillies de manière rétrospective grâce à des questionnaires standardisés, le nombre de cigarettes fumées en moyenne par jour au cours de chaque année de l'histoire d'exposition au tabagisme du sujet était disponible.

5.2 Spécification du modèle

Le modèle ZIP mixte considéré inclut une trajectoire flexible de consommation de tabac avec des fonctions splines sur le temps rétrospectif afin d'être plus représentative des trajectoires individuelles observées qui sont assez éloignées d'une trajectoire linéaire ou quadratique. Un intercept aléatoire est également considéré pour tenir compte de la corrélation entre les mesures répétées d'intensité d'un même individu. La probabilité d'être un zéro structurel est spécifique à chaque classe latente et dépend du temps avant la date index. Le modèle a été estimé pour 1 à 5 classes latentes. Pour les profils de trajectoires d'exposition identifiés, nous avons évalué leur association avec le risque de cancer du poumon en estimant un modèle de régression logistique conditionnel aux classes. Les caractéristiques des sujets classés a posteriori dans chaque classe ont enfin été décrites.

5.3 Résultats

Les résultats nous permettent d'identifier des profils de trajectoires avec des évolutions bien distinctes suivant le temps depuis la date index. Par exemple, nous avons identifié une classe avec une trajectoire moyenne constante au cours du temps ("expositions constantes modérées"). Une classe avec une trajectoire moyenne reflétant des intensités de consommation de tabac plus élevées proche de la date index a également été identifiée ("expositions élevées récentes"). Enfin, nous avons identifié une classe représentée par une trajectoire moyenne avec des intensités plus élevées plus de 30 ans avant la date index ("expositions élevées anciennes"). La classe des "expositions élevées récentes" avait le risque de cancer du poumon le plus élevé comparée à la classe des "expositions constantes modérées".

6 Conclusion

Le modèle ZIP-LCMM proposé permet d'étendre la théorie des modèles mixtes à classes latentes aux mesures répétées ayant une large proportion de zéros. Ce type de variable est particulièrement fréquent dès lors que l'on s'intéresse à des données d'expositions environnementales prolongées recueillies dans les études épidémiologiques. Il parait donc important de pouvoir fournir un modèle permettant d'identifier des profils de trajectoires tout en prenant compte la spécificité de la distribution et des trajectoires de ces expositions vie entière.

Ce développement méthodologique, motivé par l'analyse des études épidémiologiques cas-témoins, donne de nouvelles perspectives pour l'identification de trajectoires d'expositions environnementales vie entière. En couplant la richesse des données recueillies dans les études épidémiologiques à cet outil statistique, on peut considérer l'aspect temporel des

relations dose-effet, ce qui est un défi majeur en épidémiologie environnementale. Nous avons illustré l'intérêt du modèle ZIP-LCMM pour identifier des trajectoires de consommation de tabac et comparer le risque de cancer du poumon associé, mais le modèle peut aussi être appliqué pour étudier d'autres relations entre des expositions prolongées et la survenue de maladie.

Bibliographie

Ferrer L., Rondeau V., Dignam J., Pickles T., Jacqmin-Gadda H., and Proust-lima C. (2016). Joint modelling of longitudinal and multi-state process: application to clinical progressions in prostate cancer, *Statistics in Medicine*, 35(22):3933-3948.

Lambert D. (1992). Zero-Inflated Poisson Regression, with an Application to Detects in Manufacturing, *Technometrics*, 34(1):1-14.

Luce D. and Stucker I. (2011). Investigation of occupational and environmental causes of respiratory cancers (ICARE): a multicenter, population-based case-control study in France, *BMC Public Health*, 11(1):928.

Muthén B. and Shedden K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm, *Biometrics*, 55(2):463-469.

Muthén L.K and Muthén B.O. (2005), *Mplus: Statistical analysis with latent variables: User's guide*, Muthén & Muthén Los Angeles.

Proust-Lima C., Philipps V. and Liquet B. (2017). Estimation of extended mixed models using latent classes and latent processes: the R package lcmm, *Journal of Statistical Software*, 78(2):1-56.

Rizopoulos D. (2012), *Joint models for longitudinal and time-to-event data: With applications in R*, Chapman and Hall/ CRC.