Prédiction dynamique non paramétrique d'un risque d'événement à partir de prédicteurs longitudinaux

Corentin Ségalas [1,*], Cécile Proust-Lima [2] & Robin Genuer [1]

[1] *Univ. Bordeaux, INSERM, INRIA, BPH, U1219, F-33000 Bordeaux, France*
[2] *Univ. Bordeaux, INSERM, BPH, U1219, F-33000 Bordeaux, France*
[*] *corentin.segalas@u-bordeaux.fr*

Résumé. Prédire dynamiquement un risque de survenue d'évènement en prenant en compte l'historique médical complet d'un patient représente un défi statistique. En effet, de tels prédicteurs incluent souvent des variables qui évoluent au cours du temps et pour lesquelles on ne possède que des observations bruitées, mesurées à des temps irréguliers. Les approches proposées dans la littérature ont d'importantes limites. L'estimation des modèles conjoints devient impossible lorsque le nombre de prédicteurs longitudinaux croît et l'approche par regression calibration en deux étapes ignore la présence de données manquantes informatives. On propose une approche totalement non paramétrique, robuste aux données manquantes et qui permet d'inclure un grand nombre de prédicteurs longitudinaux, potentiellement mesurés irrégulièrement. Cette nouvelle méthode combine le principe des forêts aléatoires de survie (capables de gérer naturellement l'aspect grande dimension et la prédiction dynamique) avec l'analyse en composantes principales fonctionnelles (qui permet de résumer la dynamique temporelle des marqueurs).

Mots-clés. Analyse en composantes principales fonctionnelles, Données longitudinales, Forêts aléatoires, Prédiction dynamique.

Abstract. Dynamic prediction of the risk of an event taking into account a patient's complete medical history represents a statistical challenge. Indeed, such predictors often include variables which evolve over time and for which we only have noisy observations, measured at irregular times. The approaches proposed in the literature have important limitations. The estimation of joint models becomes impossible when the number of longitudinal predictors increases and the two-step regression calibration approach ignores the presence of informative missing data. We propose a completely non-parametric approach, robust to missing data and which allows the inclusion of a large number of longitudinal predictors, measured irregularly. This new method combines the principle of survival random forests (capable of naturally managing the high dimension aspect and dynamic prediction) with functional principal component analysis (which allows the temporal dynamics of the markers to be taken into account).

Keywords. Functional principal component analysis, Longitudinal data, Non parametric dynamic prediction, Random forests.

1 Introduction

En santé, il est courant de vouloir prédire le risque individuel de survenue d'un événement à partir de l'historique médical d'un patient. Implémenter de tels modèles représente un challenge statistique car ils doivent pouvoir incorporer un grand nombre de prédicteurs, dont certains évoluent au cours du temps. De plus, dans les études de santé, ces prédicteurs sont souvent observés de manière irrégulière, avec erreur de mesure et leur trajectoire temporelle peut être tronquée par la survenue de l'évènement. Les méthodes classiques, basées sur le modèle de Cox par exemple, ne sont adaptées à la prise en compte de tels prédicteurs.

Dans la littérature, trois grandes approches ont été proposées pour prédire un risque d'événement à partir de données longitudinales :

- l'approche par regression calibration [1] qui modélise séparément les trajectoires des prédicteurs puis les inclut dans des modèles de prédiction. Le problème de cette approche est que ce fonctionnement en deux étapes présente un biais en cas de données manquantes informatives.
- l'approche landmark [10] qui se place à un temps de prédiction et utilise tout l'historique jusqu'à ce temps pour prédire la survenue ultérieure de l'évènement. Ici, la censure informative est prise en compte mais au prix d'une réduction de l'information disponible.
- l'approche par modèle conjoint [8] qui modélise simultanément les trajectoires des prédicteurs et le risque d'évènement. Mais l'estimation des modèles conjoints devient trop lourde quand le nombre de prédicteurs longitudinaux croît.

Les forêts aléatoires [2] constituent un modèle prédictif dont l'avantage est de pouvoir modéliser des relations complexes et non linéaires entre prédicteurs en agrégeant un ensemble d'arbres de décision. Un arbre effectue un partitionnement récursif binaire de l'espace des prédicteurs en régions de plus en plus homogènes (en terme de variable à prédire). Elles ont été étendues au contexte de l'analyse de survie [5] mais sans pouvoir inclure de prédicteurs longitudinaux. Dans un précédent travail [3], les forêts dynamiques de survie ont été développées. Au sein de chaque noeud et pour chaque arbre de la forêt, chaque prédicteur longitudinal sélectionné comme candidat potentiel pour le calcul de la séparation optimale est modélisé par un modèle mixte [6], puis résumé à l'aide des effets aléatoires individuels prédits. De fait, cette approche est paramétrique et nécessite de spécifier les modèles mixtes. Cela peut impacter le temps de calcul du modèle et ne permet pas d'envisager des prédicteurs longitudinaux mesurés de façon intensive. Dans ce travail, on étend la méthode des forêts dynamiques en résumant les prédicteurs longitudinaux par les scores individuels issus d'une analyse en composante principale fonctionnelle [7] conduisant à une approche complètement non-paramétrique.

2 Méthode

Pour chaque participant $i \in \{1, \ldots, N\}$, on note T_i le temps de survenue de l'évènement, C_i le temps de censure supposé indépendant de T_i et l'on observe alors $\tilde{T}_i = \min(T_i, C_i)$. On note δ_i l'indicateur d'évènement qui vaut $k \in \{1, \ldots, K\}$ si l'évènement de cause k est

observé, 0 s'il est censuré. On collecte également P prédicteurs indépendants du temps, X_{ip} avec $p \in \{1, \ldots, P\}$ et Q prédicteurs dépendants du temps Y_{ijq} avec $j \in \{1, \ldots, n_{iq}\}$ et $q \in \{1, \ldots, Q\}$ mesurés aux temps $t_{ijm} \leq \tilde{T}_i$.

Le principe de la forêt aléatoire est le suivant : la prédiction finale s'obtient en agrégeant les prédictions d'un ensemble d'arbres de décision. Chaque arbre est construit sur un échantillon bootstrap de l'échantillon initial et procède par partitionnement binaire récursif de l'espace des prédicteurs. La séparation choisie est celle qui maximise la distance (en terme de variable à prédire) entre les deux sous-groupes. Les forêts aléatoires incluent un aléa supplémentaire en ne considérant comme candidats à la division qu'une sélection aléatoire de ces prédicteurs. La méthode des forêts aléatoires a été étendue à l'analyse de survie en utilisant comme distance la statistique du logrank [5] et en proposant une méthode d'agrégation des arbres adaptée. La méthode est aussi applicable dans un contexte de risques compétitifs en considérant le test de Fine and Gray [4].

Néanmoins, ces dernières ne permettent pas d'inclure des prédicteurs dépendant du temps. C'est pourquoi les forêts dynamiques [3] ont été introduites. Le principe est le suivant : pour chacun des arbres de la forêt, au sein de chaque noeud et avant toute division, une fois la sélection aléatoires des prédicteurs effectuée, les trajectoires individuelles des prédicteurs longitudinaux (Y_{ijq}) vont être résumées par des quantités rendues indépendantes du temps : les effets aléatoires prédits après estimation d'un modèles mixte sur $(Y_{ijq})_{ij}$.

Dans ce travail, on conserve la même architecture de construction de la forêt aléatoire, mais en utilisant les scores issus de l'analyse en composante principale fonctionnelle (ACPF) pour résumer les trajectoires longitudinales plutôt que les effets aléatoires prédits par les modèles mixtes. L'avantage de cette approche est qu'elle évite de faire des hypothèses paramétriques sur les formes des trajectoires ou sur la distribution des effets aléatoires contrairement à ce qui a été proposé [3]. Si l'ACPF a été initialement développée pour des données fonctionnelles denses et régulières [7], l'algorithme PACE [11] a été introduit pour adapter l'ACPF à des données fonctionnelles éparses et irrégulières.

Plaçons nous dans le cadre de l'analyse de données fonctionnelles et supposons que, pour un $q \in \{1, \ldots, Q\}$ fixé, les trajectoires $(Y_{ijq})_{ij}$ sont la collection bruitée de réalisations aléatoires $(f_{iq})_i$ d'une fonction inconnue sous-jacente f_q. L'ACPF est basée sur la décomposition de Karhunen-Loève qui, sous des hypothèses de régularité, assure que

$$f_{iq}(t) = \mu_q(t) + \sum_{k=1}^{\infty} \xi_{iqk}\phi_{qk}(t)$$

où ξ_{ikq} et $\phi_{qk}(t)$ sont respectivement les valeurs propres (aussi appelés scores de l'ACPF) et les fonctions propres issues de la décomposition en valeur propre de l'opérateur de covariance de f_q. Afin de réduire la dimension infinie de cette décomposition, on peut ne conserver que les K premiers termes de la somme et les scores ξ_{iqk} de l'ACPF représentent alors les coordonnées dans cet espace fonctionnel de dimension fini. Pour un individu i, ils mesurent la déviation individuelle à la fonction moyenne $\mu_q(t)$ et constituent un résumé intéressant de sa dynamique temporelle. K peut être fixé arbitrairement ou en se basant sur des critères de pourcentage de variance expliquée. L'algorithme PACE [11] permet directement d'obtenir des estimations de ces scores $\hat{\xi}_{iqk}$ mais aussi de la fonction moyenne $\hat{\mu}_q(\mathbf{t})$ et des composantes

principales fonctionnelles $\hat{\phi}_{qk}(\mathbf{t})$ sur une grille de temps $\mathbf{t}$.

Ainsi, au sein de chaque noeud et avant toute division, une fois les prédicteurs candidats sélectionnés (parmi les P indépendants du temps et les Q dépendants du temps), on applique aux candidats longitudinaux l'algorithme PACE. Pour chaque participant $i \in \{1, \ldots, N\}$, on ne disposera alors que de prédicteurs indépendants du temps (des variables indépendantes du temps et des scores estimés par l'ACPF $\hat{\xi}_{iqk}$). On peut maintenant appliquer la stratégie usuelle des forêts aléatoires de survie [5] puis effectuer la division optimale avec pour critère la statistique du log-rank. La construction de l'arbre se termine lorsque le critère d'arrêt est atteint (le nombre d'évènements minimal dans un noeud est atteint par exemple). À ce stade, on suppose que les noeuds terminaux sont suffisamment homogènes en terme de probabilité de survenue de l'évènement. Dans chacun de ces noeuds terminaux, la fonction de risque cumulé est estimée par l'estimateur de Nelson-Aalen.

Pour un nouvel individu, il est possible de faire de la prédiction dynamique de survenue de l'évènement à partir d'un temps de prédiction s pour un horizon donné. Pour cela, il faut tout d'abord tronquer pour cet individu les prédicteurs longitudinaux au temps s. Puis on peut prédire avec la forêt l'évènement à partir de ces données. Cela nécessite, au sein de chaque noeud de calculer les coordonnées de cet individu dans la base de l'espace fonctionnel de dimension K obtenue sur l'échantillon d'apprentissage. L'estimation finale s'obtient alors en agrégeant les probabilités prédites de l'évènement dans les feuilles où le sujet est classé.

3 Simulation et application

À la différence des modèles mixtes, connus pour être robustes aux données manquantes dites *missing at random* (la probabilité qu'un marqueur ne soit pas observé ne dépend pas des valeurs non observées de ce marqueur), le comportement de l'ACPF en présence de données manquantes est incertain. Nous avons mené une étude de simulation préliminaire [9] qui a permis d'assurer que l'ACPF, sous divers scénarios de données manquantes, se comportait aussi bien qu'un modèle mixte à l'exception des scénarios les plus extrêmes. Cet argument, associé au fait que, par l'architecture de la forêt aléatoire de survie, les profils de survie des individus tendent à se ressembler, vont dans le sens d'une bonne robustesse aux données manquantes *missing at random*.

Nous avons réalisé une étude de simulation pour valider cette nouvelle approche en étudiant l'impact des données manquantes et de la stratégie de choix de K. On s'est aussi intéressé à la prise en compte de la variabilité de la trajectoire temporelle. Pour cela, deux stratégies ont été envisagées pour résumer l'information longitudinale : ACPF sur la trajectoire uniquement, ACPF sur la trajectoire et sa dérivée. Enfin, nous avons comparé cette approche à l'approche des forêts dynamiques par modèles mixtes proposée précédemment [3]. Nous présenterons les résultats de cette étude de simulation.

Nous présenterons également une application sur données réelles. L'objectif était de prédire le risque de survenue de vasospasme chez les patients hospitalisés après une hémorragie sub-arachnoïdienne (HSA). Il s'agit d'une complication majeure qui a lieu entre 3 et 14 jours après la survenue de la HSA, souvent détecté trop tard pour un traitement efficace. Dans les

facteurs de risque identifiés, certains sont monitorés au cours de l'hospitalisation de ces patients (par exemple la pression artérielle, l'hyperglycémie), toutes les heures pour la plupart. Ainsi l'approche des forêts dynamiques que nous avons développée est adaptée à ce genre de contexte et permettrait de construire un modèle prédictif utile pour aider le clinicien dans le suivi des patients.

4 Conclusion

Nous avons étendu la méthode des forêts dynamiques [3] et proposé une nouvelle approche totalement non paramétrique qui permet de faire de la prédiction dynamique du risque de survenue d'un évènement en incluant des prédicteurs, potentiellement en grand nombre, dépendant du temps, mesurés irrégulièrement et qui peuvent inclure des données manquantes. De plus, cette approche, basée sur la méthode des forêts aléatoires de survie, bénéficie d'outils statistiques qui permettent d'informer et d'évaluer l'importance des variables dans la prédiction. Cela permet d'éclairer les mécanismes de prédiction du modèle, évitant l'écueil de l'effet *boîte noire* de certains algorithmes prédictifs.

Références

[1] Paul S. Albert et Joanna H. Shih : On Estimating the Relationship between Longitudinal Measurements and Time-to-Event Data Using a Simple Two-Stage Procedure. *Biometrics*, 66(3):983–987, 2010.

[2] Leo Breiman : Random Forests. *Machine Learning*, 45(1):5–32, octobre 2001.

[3] Anthony Devaux, Catherine Helmer, Robin Genuer et Cécile Proust-Lima : Random survival forests with multivariate longitudinal endogenous covariates. *Statistical Methods in Medical Research*, octobre 2023.

[4] Hemant Ishwaran, Thomas A. Gerds, Udaya B. Kogalur, Richard D. Moore, Stephen J. Gange et Bryan M. Lau : Random survival forests for competing risks. *Biostatistics (Oxford, England)*, 15(4):757–773, octobre 2014.

[5] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone et Michael S. Lauer : Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, septembre 2008.

[6] Geert Molenberghs et Geert Verbeke : *Linear Mixed Models for Longitudinal Data*. Springer Series in Statistics. Springer, New York, NY, 2000.

[7] James O. Ramsay et Bernard W. Silverman : *Functional Data Analysis*. Springer Series in Statistics. Springer, New York, NY, 2005.

[8] Dimitris Rizopoulos : *Joint Models for Longitudinal and Time-to-Event Data : With Applications in R*. CRC Press, juin 2012.

[9] Corentin Ségalas, Catherine Helmer, Robin Genuer et Cécile Proust-Lima : Functional principal component analysis as an alternative to mixed-effect models for describing sparse repeated measures in presence of missing data, février 2024. arXiv :2402.10624 [stat].

[10] Hans C. Van Houwelingen : Dynamic Prediction by Landmarking in Event History Analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007.

[11] Fang Yao, Hans-Georg Müller et Jane-Ling Wang : Functional Data Analysis for Sparse Longitudinal Data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.