Splines multidimensionnelles pénalisées dans les modèles de survie : applications en épidémiologie des cancers

Mathieu FAUVERNIER, Laurent ROCHE, Zoé UHRY, Laure TRON, Nadine BOSSARD, Laurent REMONTET

EPICLIN 2019, TOULOUSE

Vendredi 17 Mai 2019

Contexte épidémiologique

Le **cancer** est la **première cause de mortalité** en France (30% des décès)

Pour un cancer donné, connaître les **facteurs pronostiques** de la **survie** est un enjeu majeur

Facteurs pronostiques = caractéristiques

- du patient
- de la maladie
- de la prise en charge
- de l'**environnement social et économique**

Contexte méthodologique

Survie nette : survie que l'on observerait si la seule cause de décès était le cancer étudié.

Au temps t, le taux de mortalité toutes causes $h_{ttc}(t)$ est décomposé ainsi :

$$h_{ttc}(t|\boldsymbol{X}, \boldsymbol{Z}) = h_{attendu}(t|\boldsymbol{Z}) + h_{excès}(t|\boldsymbol{X})$$

Avec $\boldsymbol{X}$ les covariables d'intérêt et $\boldsymbol{Z}$ les covariables démographiques.

Modélisation du taux en excès

$$Log[h_{excès}(t|\boldsymbol{X})] = f(t, \boldsymbol{X})$$

$\boldsymbol{X}$ peut contenir : âge au diagnostic, année de diagnostic, indice de défavorisation socio-économique, stade, ...

Contexte méthodologique

Effet non linéaire : effet d'un an d'âge supplémentaire dépend de l'âge considéré

Effet non proportionnel : effet de l'âge dépend du temps de suivi considéré

Interaction : effet de l'année dépend de l'âge considéré

Modélisation complexe

Pour représenter f, il faut tenir compte simultanément des effets potentiellement non linéaires et non proportionnels des effets propres de chaque covariable, mais également des interactions !

Objectif

Proposer un modèle du taux de mortalité (en excès) à partir de splines multidimensionnelles pénalisées

Splines = polynômes par morceaux utilisés pour représenter des effets non linéaires

Cadre théorique développé par Wood et al. (2016) mais adaptation nécessaire aux modèles de taux en excès

Méthode - spline multidimensionnelle par produit tensoriel

Considérons un effet linéaire de l'âge (pour simplifier) :

$$f(a) = \alpha_0 + \alpha_1 \times a$$

Faisons varier l'intercept et la pente selon un effet quadratique g du temps : $g(t) = \gamma_0 + \gamma_1 \times t + \gamma_2 \times t^2$.

Méthode - spline multidimensionnelle par produit tensoriel

Considérons un effet linéaire de l'âge (pour simplifier) :

$$f(a) = \alpha_0 + \alpha_1 \times a$$

Faisons varier l'intercept et la pente selon un effet quadratique g du temps : $g(t) = \gamma_0 + \gamma_1 \times t + \gamma_2 \times t^2$.

L'effet bidimensionnel est alors donné par :

$$\begin{aligned} tensor(t, age) &= \beta_0 + \beta_1 \times t + \beta_2 \times t^2 && (1) \\ &+ (\beta_3 + \beta_4 \times t + \beta_5 \times t^2) \times a && (2) \end{aligned}$$

Méthode - spline multidimensionnelle par produit tensoriel

Considérons un effet linéaire de l'âge (pour simplifier) :

$$f(a) = \alpha_0 + \alpha_1 \times a$$

Faisons varier l'intercept et la pente selon un effet quadratique g du temps : $g(t) = \gamma_0 + \gamma_1 \times t + \gamma_2 \times t^2$.

L'effet bidimensionnel est alors donné par :

$$tensor(t, age) = \beta_0 + \beta_1 \times t + \beta_2 \times t^2 \quad (1)$$

$$+ (\beta_3 + \beta_4 \times t + \beta_5 \times t^2) \times a \quad (2)$$

⇒ Prise en compte de l'interaction temps*âge

Méthode - spline multidimensionnelle par produit tensoriel

Considérons un effet linéaire de l'âge (pour simplifier) :

$$f(a) = \alpha_0 + \alpha_1 \times a$$

Faisons varier l'intercept et la pente selon un effet quadratique g du temps : $g(t) = \gamma_0 + \gamma_1 \times t + \gamma_2 \times t^2$.

L'effet bidimensionnel est alors donné par :

$$\begin{aligned} tensor(t, age) &= \beta_0 + \beta_1 \times t + \beta_2 \times t^2 & (1) \\ &+ (\beta_3 + \beta_4 \times t + \beta_5 \times t^2) \times a & (2) \end{aligned}$$

$\Rightarrow$ Prise en compte de l'interaction temps*âge

Le principe est similaire avec plus de deux covariables.

Méthode - pénalisation

En règle générale, on n'utilise pas des polynômes linéaires ou quadratiques mais des splines cubiques.

Problèmes

- Le nombre de paramètres augmente fortement avec le nombre de facteurs pronostiques et la complexité des splines
- Les prédictions risquent d'être erratiques

Solution proposée

Pénalisation visant à lisser les prédictions

Méthode - vraisemblance pénalisée

On considère le temps comme seul prédicteur avec

$$log[h_{\text{excès}}(t)] = f(t)$$

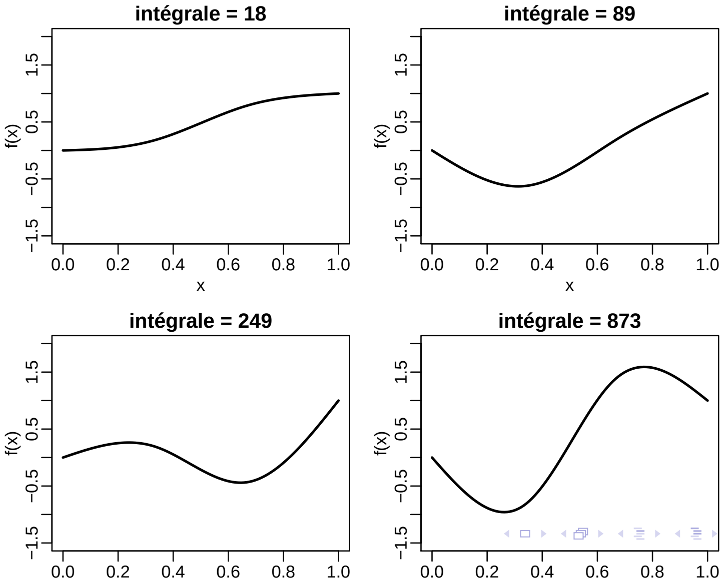
Si $l(\beta)$ est la log-vraisemblance du modèle et λ le paramètre de lissage, alors la log-vraisemblance pénalisée $l_p(\beta|\lambda)$ s'écrit :

$$l_p(\beta|\lambda) = l(\beta) - \lambda pen = l(\beta) - \lambda \int f''(t)^2 \mathrm{d}t$$

Le critère objectif $l_p(\beta|\lambda)$ est un compromis entre la fidélité aux données $l(\beta)$ et la régularité de l'effet du temps $\int f''(t)^2 \mathrm{d}t$.

λ est le paramètre qui contrôle ce compromis.

Pénalisation sur la dérivée seconde (pourquoi $\int f''(t)^2 \mathrm{d}t$?)



Méthode - Estimation des paramètres de lissage

En pratique, **il y a autant de paramètres de lissage que de facteurs pronostiques** et il faut les estimer.

Deux approches sont possibles :

- **validation croisée** : cherche le modèle qui **minimise les erreurs de prédictions**
- **vraisemblance marginale** : approche bayésienne cherchant le modèle **le plus susceptible d'avoir généré les données**

Inégalités socio-économiques

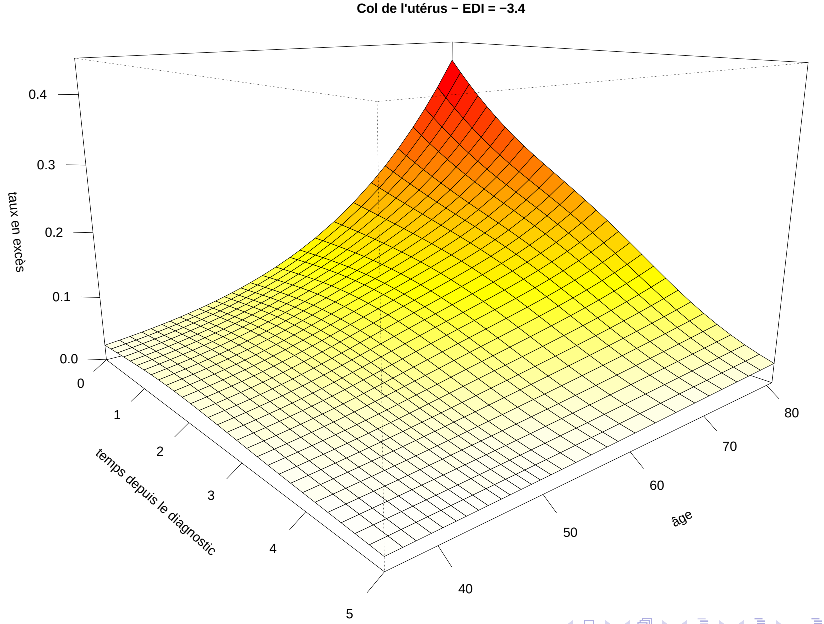
Données: 1 865 patientes diagnostiquées d'un cancer du col de l'utérus entre 2006 et 2009 en France (données FRANCIM).

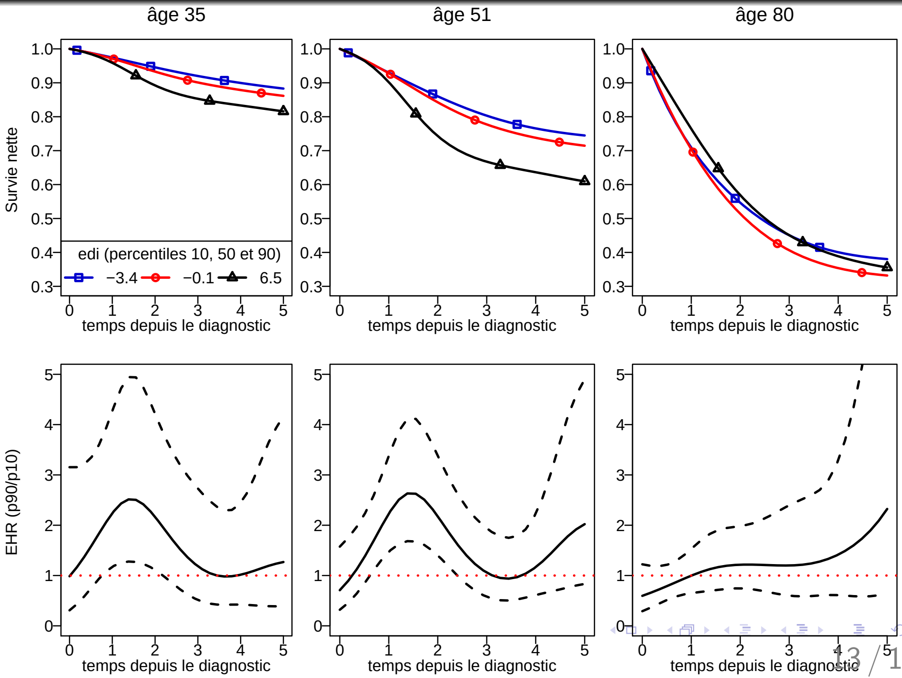
EDI = European Deprivation Index (variable continue, **valeur élevée = défavorisation élevée**)

Modèle:

$$Log[h_E(t, a, EDI)] = tensor(t, a, EDI)$$

- 5 nœuds par base marginale (temps, âge, EDI)
- Nombre de paramètres: 5*5*5 = 125
- 3 paramètres de lissage


Col de l'utérus − EDI = −3.4
taux en excès
0.4
0.3
0.2
0.1
0.0
temps depuis le diagnostic
0
1
2
3
4
5
âge
40
50
60
70
80


âge 35
âge 51
âge 80
Survie nette
EHR (p90/p10)
edi (percentiles 10, 50 et 90)
−3.4
−0.1
6.5
temps depuis le diagnostic
temps depuis le diagnostic
temps depuis le diagnostic
temps depuis le diagnostic
temps depuis le diagnostic
temps depuis le diagnostic

Conclusion

Avantages

- Modélisation simultanée des effets non linéaires, non proportionnels et des interactions
- Restitution de formes très complexes en limitant : sélection de modèles, catégorisation de variables continues et sur-ajustement
- Sélection intégrée des paramètres de lissage stable et efficiente
- Forme entièrement paramétrique permettant la projection
- Cadre théorique commun (développé par Wood) pour les vraisemblances régulières
- **Nombreuses perspectives en recherche clinique**

Inconvénients

- Approche tensor limitée à 3 ou 4 covariables continues

La méthode est décrite par Fauvernier et al. (2019) et est implémentée dans le package R **survPen** disponible sur le CRAN et sur GitHub

Remerciements

Merci de votre attention

Nous tenons à remercier l'ANR (Agence Nationale de la Recherche) et l'Institut National du Cancer (INCa) pour avoir financé ce travail. Nous remercions également Guy Launoy, Jacques Estève et le réseau FRANCIM.

Références

Fauvernier, M., Roche, L., Uhry, Z., Tron, L., Bossard, N., Remontet, L., and the CENSUR working survival group (2019). Multidimensional penalized hazard model with continuous covariates: applications for studying trends and social inequalities in cancer survival. *in revision in the Journal of the Royal Statistical Society Series C Applied Statistics*.

Wood, S. N., Pya, N., and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111(516):1548–1563.