AMÉLIORER LA SÉLECTION ET LA PRÉCISION DES MODÈLES DE SURVIE : UNE APPROCHE BAYÉSIENNE SPIKE AND SLAB

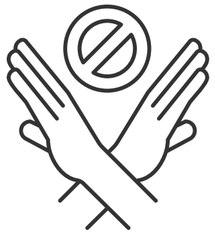
Université de Lyon | Ecole doctorale E2M2

Présenté par Joanna Pautonnier (Doctorante 1ère année)
Encadrement de thèse : Pr Pascal Roy (60%) / Pr David Causeur (40%)
Financement : PEPR Santé Numérique
12 décembre 2025

LBBE
LABORATOIRE DE BIOMÉTRIE ET BIOLOGIE ÉVOLUTIVE
IRMAR

Introduction

Les modèles de survie prédictifs sont de plus en plus utilisés en médecine de précision.



Limites actuelles : absence d'indicateurs fiables d'incertitude quand le modèle est trop complexe (GAM pénalisés).

Objectif première partie de thèse : Évaluer et améliorer les propriétés prédictives globales et contextualisées de modèles de survie et leurs incertitudes.

Méthode

Qu'est-ce que l'analyse de données de durée de survie ?

Etude des délais de la survenu d'un évènement.

Méthode

Quelle utilité ?

- Etudes longitudinale
- Essais thérapeutique

LES INDICATEURS EN SURVIE

FONCTION DE SURVIE

Probabilité de survivre jusqu'à un temps t.

$$S(t) = P(T_i > t)$$

LES INDICATEURS EN SURVIE

FONCTION DE SURVIE

Probabilité de survivre jusqu'à un temps t.

$$S(t) = P(T_i > t)$$

DENSITÉ DE PROBABILITÉ

Probabilité de mourir dans un petit intervalle de temps entre t et t+dt, à t fixé.

$$f(t) = \lim_{dt \to 0} \frac{P(t \leq T_i \leq t + dt)}{dt}$$

LES INDICATEURS EN SURVIE

FONCTION DE SURVIE

Probabilité de survivre jusqu'à un temps t.

$$S(t) = P(T_i > t)$$

DENSITÉ DE PROBABILITÉ

Probabilité de mourir dans un petit intervalle de temps entre t et t+dt.

$$f(t) = \lim_{dt \to 0} \frac{P(t \leq T_i \leq t + dt)}{dt}$$

TAUX DE HAZARD

Densité conditionnelle au fait d'avoir survécue jusqu'à ce temps.

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T_i \leq t + dt \mid T_i > t)}{dt} = \frac{f(t)}{S(t)}$$

LES INDICATEURS EN SURVIE

$$\frac{\text{Probabilité de mourir dans un petit intervalle après t sachant qu'on a survécu jusqu'à t.}}{\text{Petit intervalle de temps}} = \frac{\text{sans unité}}{\text{en unité de temps (jours, mois ect.)}}$$

Le taux de hazard est une vitesse/force d'évènement instantanée et s'exprime en évènement par temps.

TAUX DE HAZARD

Densité conditionnelle au fait d'avoir survécue jusqu'à ce temps.

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T_i \leq t + dt \mid T_i > t)}{dt} = \frac{f(t)}{S(t)}$$

LES INDICATEURS EN SURVIE

Survie en fonction du taux de hazard :

$$S(t) = exp(-\int_0^t h(u)\,\mathrm{d}u)$$

Hazard

- Vision : la force de mortalité à chaque instant t.
- L'histoire racontée : révèle l'évolution du risque.
- Pour le clinicien : comprendre le "Pourquoi/Quand".

Survie

- Vision : la probabilité de n'avoir pas subi l'événement jusqu'à t.
- L'histoire racontée : lisse les variations pour donner un bilan global.
- Pour le clinicien : faire la prédiction individuelle (Prognostic).

DU MODÈLE DE COX À LA MODÉLISATION FLEXIBLE PAR SPLINES

Modèle semi-paramétrique sur le risque instantané :

$$h(t|X) = h_0(t)exp(X\beta)$$

Taux de base : dépend uniquement du temps.

Dépend uniquement des patients, pas du temps :

- β : vecteur de paramètres (à estimer)
- X: vecteur de covariables (mesuré)

Hypothèse de proportionnalité :

- β ne dépend pas du temps = l'effet est constant dans le temps

➡ **SOUVENT FAUSSE**.

DU MODÈLE DE COX À LA MODÉLISATION FLEXIBLE PAR SPLINES

Non estimation du taux de base

- $h_0(t)$ **:** difficile à estimer.
- Le rapport des hazards entre individu ne dépend pas de $h_0(t)$: suffisant pour estimer les effets

$$\frac{h\left(t\,|X_i\right)}{h\left(t\,|X_j\right)} = exp\left(\beta\left(X_i - X_j\right)\right)$$

- **Vraisemblance partielle (ne dépends que des rapports)** : Sachant qu'un décès s'est produit à l'instant ti parmi le groupe, quelle est la probabilité que ce soit l'individu i plutôt qu'un autre ?

risque du patient i

$$L_i\left(\beta\right) = \frac{h_0\left(t_i\right)exp\left(X_i\beta\right)}{\sum_{j\in R(t_i)} h_0\left(t_i\right)exp\left(X_j\beta\right)} = \frac{exp\left(X_i\beta\right)}{\sum_{j\in R(t_i)} exp\left(X_j\beta\right)}$$

somme des risques de tous les patients présents à cet instant

Sans taux de base estimé

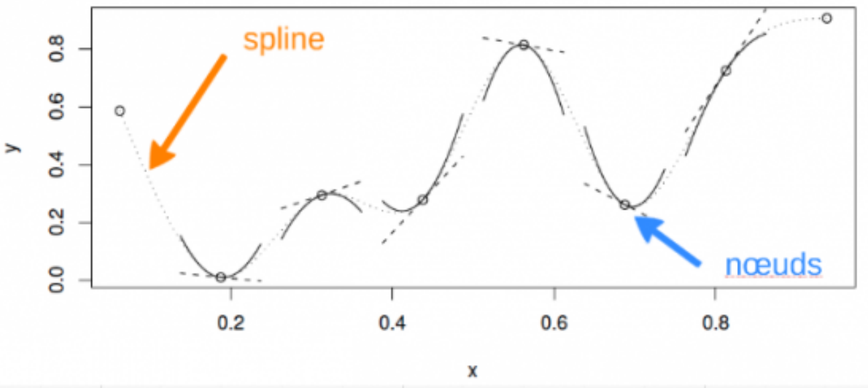
- **pas de prédictions individuelles possible.**
- **pas de compréhension de l'evolution du risque dans le temps**

DU MODÈLE DE COX À **LA MODÉLISATION FLEXIBLE PAR SPLINES**

Modélisation flexible par splines + vraisemblance totale

1. **Le risque de base** n'est plus un paramètre de nuisance mais une fonction lisse estimée (f(t)).
2. **L'effet des variables** : peut être fonction du temps (interaction X * temps).

$$\ln\left(h\left(t\,|X\right)\right) = f\left(t\right)\sum g_j\left(t\right)X_j$$



Spline : Fonctions polynomiales définies par morceaux

Avantages :

- **Taux de base est estimé**
- **Pas d'hypothèse de proportionnalité**

CONSTRUIRE UN MODÈLE : ÉTAPES CLÉS

Sélection de variables

Sélection de formes fonctionnelles

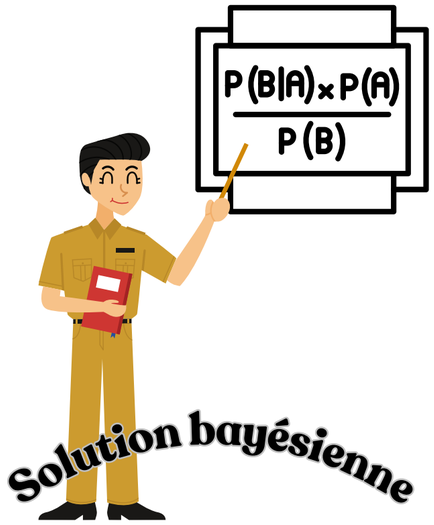
Sélection d'interaction

PROBLEMATIQUE :

ACTUELLEMENT EN FRÉQUENTISTE

- Méthodes de sélection : couteux en temps
- Intervalle de **CONFIANCE** présent mais pas exacte pour les modèles complexes.
- Absence d'intervalle de prédiction généralement.

Introduction



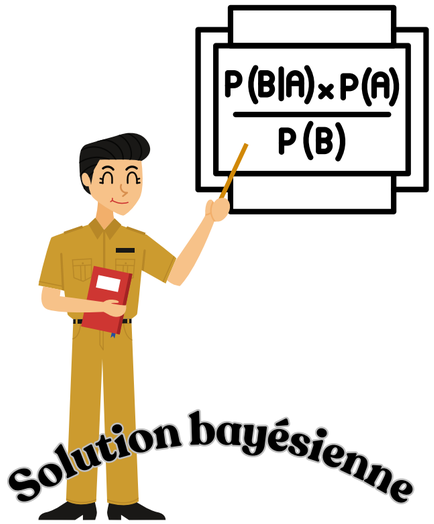
Incertitude — Accès aux distributions a posteriori.

Séléction — Prior (**spike and slab**, Laplace, gaussien…) : sélection automatique des variables + contrôle de la complexité.

Implémentation — Modèle de COX : Package R brms, BayesSurvive, psbcGroup, BMA..

Problématique — Hypothèse de proportionnalité souvent fausses + taux de base non modélisé

Introduction

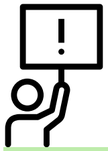


Implémentation

Approximation du Hazard par Poisson piecewise : Package R **SpikeSlabGAM** de F. Scheipl (2011)

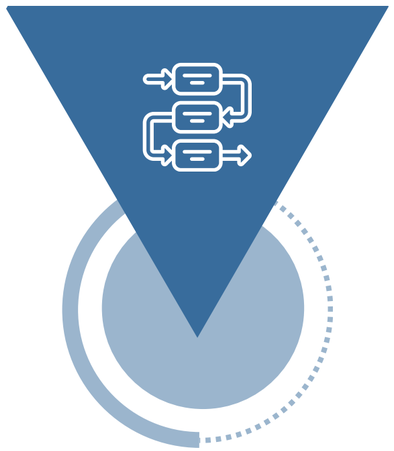
Problématique

Approximation de poisson : Duplication massive des données. => Problème computationnel



Objectif : développer une alternative bayésienne directe de SpikeSlabGAM pour les modèles de taux.

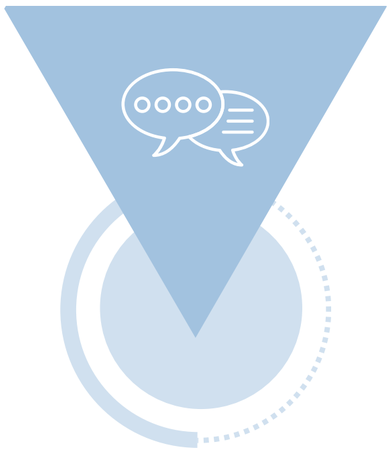
OBJECTIF DE LA PRESENTATION



Cadre méthodologique

Implémentation proposée

Résultats préliminaires obtenus par simulation



Discussion et perspectives pour la suite

EQUIVALENCE **LOG-VRAISEMBLANCE DU TAUX** VS APPROXIMATION DE POISSON

Pour un individu i : $$L_i(\beta) = \delta_i \log\left(h(t_i)\right) - \int_0^{t_i} h(u)\,du$$

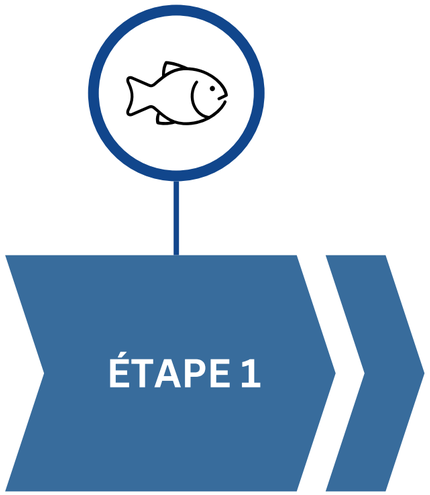
Indicateur d'évènement.

Difficulté : l'intégrale n'a souvent pas de solution analytique.

⇨ Approximation numérique : Gauss-Legendre.

Globale : somme des contributions individuelles

EQUIVALENCE LOG-VRAISEMBLANCE DU TAUX VS **APPROXIMATION DE POISSON**



Vraisemblance de poisson

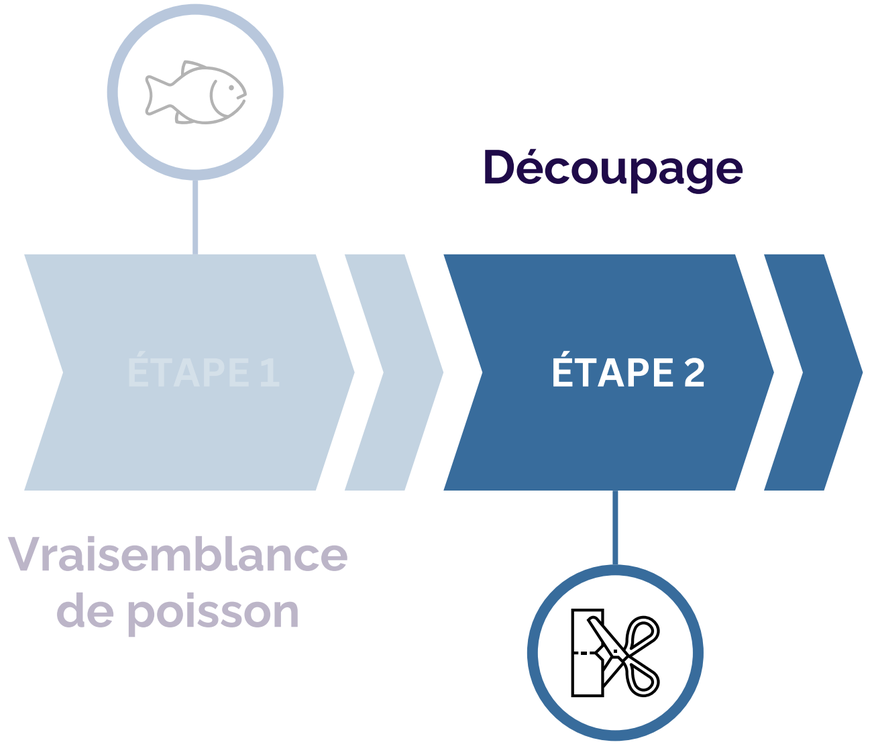
On suppose que δ_i suit une loi de poisson $\delta_i \sim \mathcal{P}(\mu_i)$
La vraisemblance du modèle est :

$$l^{\text{Poisson}}(\beta) = \prod_{i=1}^{n} \frac{\mu_i^{\delta_i} e^{-\mu_i}}{\delta_i!}$$

log-vraisemblance

$$L^{\text{Poisson}}(\beta) = \sum_{i=1}^{n} \left[\delta_i \log(\mu_i) - \log(\delta_i!) - \mu_i \right]$$

EQUIVALENCE LOG-VRAISEMBLANCE DU TAUX VS **APPROXIMATION DE POISSON**



On découpe nos données en m intervalles de temps $[d_j; d_{j+1}]$
La log-vraisemblance pour un individu i à un intervalle de temps j est :

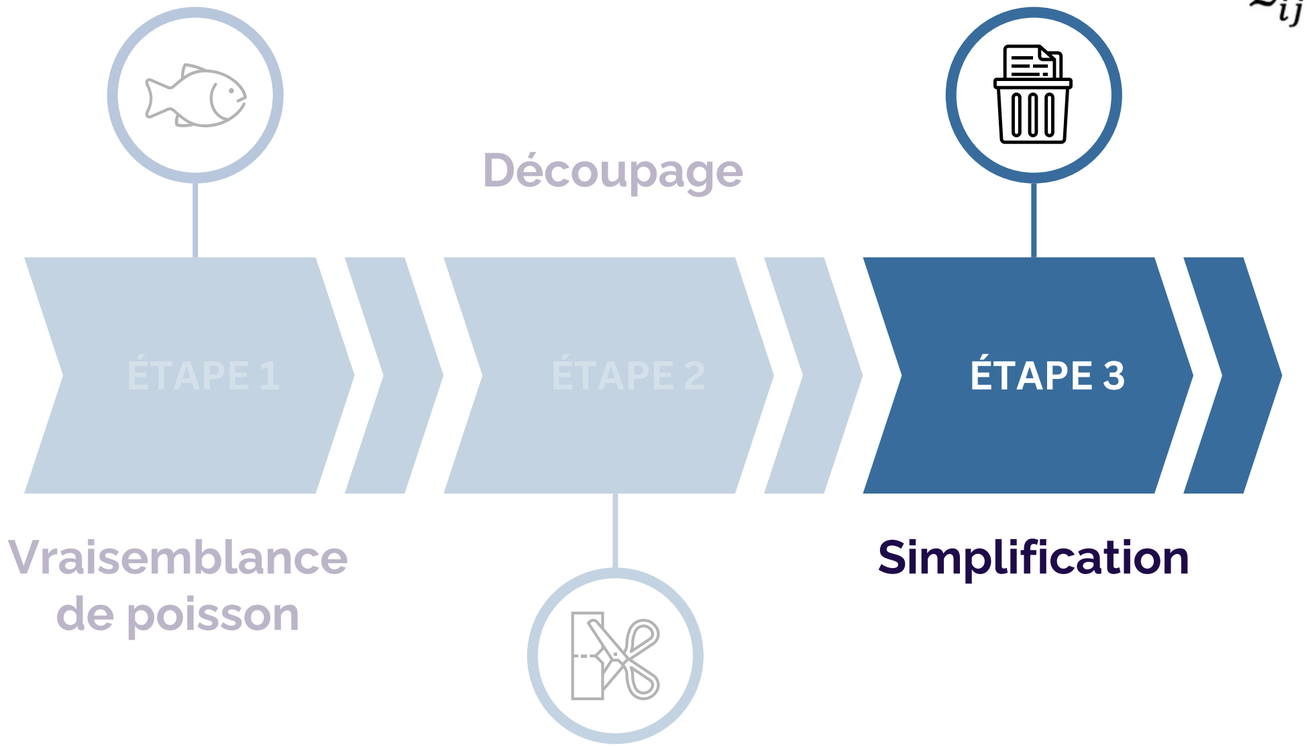
$$\mathcal{L}_{ij}^{poisson}(\beta) = \delta_{ij}\log(\mu_{ij}) - \log(\delta_{ij}!) - \mu_{ij}$$

$\mu_{ij} = (d_{j+1} - d_j)h(r_{ij}; x_i)$,

- $(d_{j+1} - d_j)$: durée de l'intervalle
- $h(r_{ij}; x_i)$: le taux calculé à un point milieu

$$\mathcal{L}_{ij}^{poisson}(\beta) = \delta_{ij}\log\left(h(r_{ij}; x_i)\right) + \delta_{ij}\log\left((d_{j+1} - d_j)\right) - \log(\delta_{ij}!) - (d_{j+1} - d_j)h(r_{ij}; x_i)$$

EQUIVALENCE LOG-VRAISEMBLANCE DU TAUX VS **APPROXIMATION DE POISSON**



$$\mathcal{L}_{ij}^{poisson}(\beta) = \delta_{ij} \log\Big(h(r_{ij}; x_i)\Big) + \delta_{ij} \log\Big((d_{j+1} - d_j)\Big) - \log(\delta_{ij}!) - (d_{j+1} - d_j) h(r_{ij}; x_i)$$

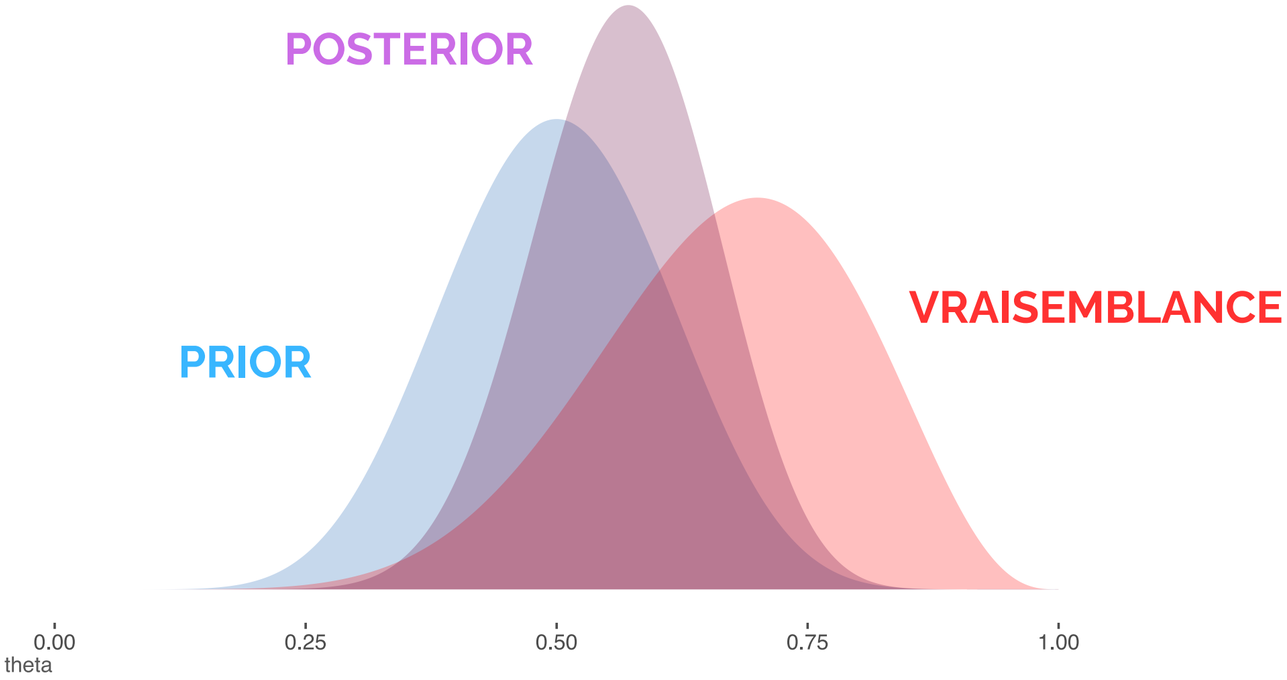
Suppression des termes non liés à beta.

$$\mathcal{L}_{ij}^{poisson}(\beta) = \delta_{ij} \log\Big(h(r_{ij}; x_i)\Big) - (d_{j+1} - d_j) h(r_{ij}; x_i)$$

où $(d_{j+1} - d_j) h(r_{ij}; x_i) \approx \int_d^{d+1} h(u; x_i) du$

≈ log-vraisemblance du taux

Méthode



$$P(\beta|données) = \frac{P(données|\beta) \times P(\beta)}{P(données)}$$

$$P(\beta|données) \propto P(données|\beta) \times P(\beta)$$

$$posterior \propto vraisemblance \times prior$$

PRIOR

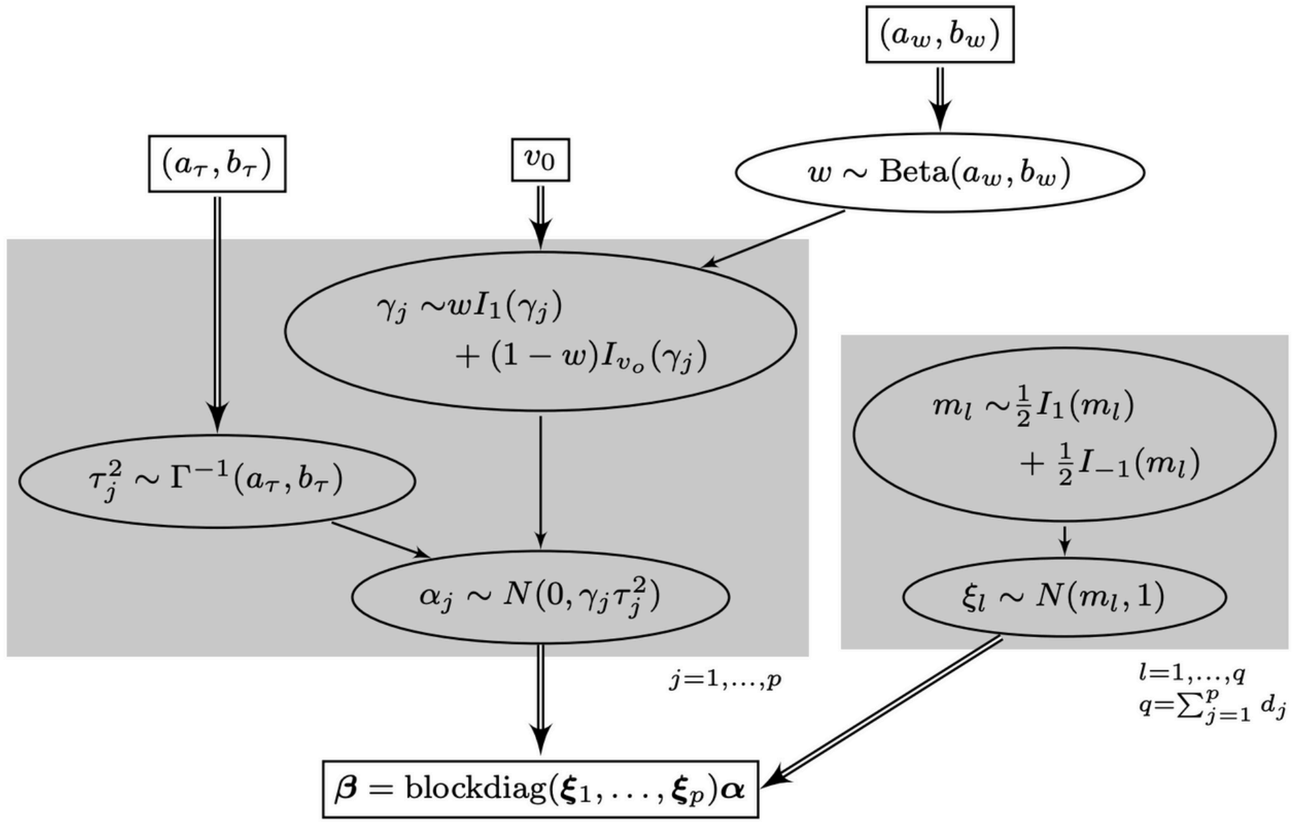
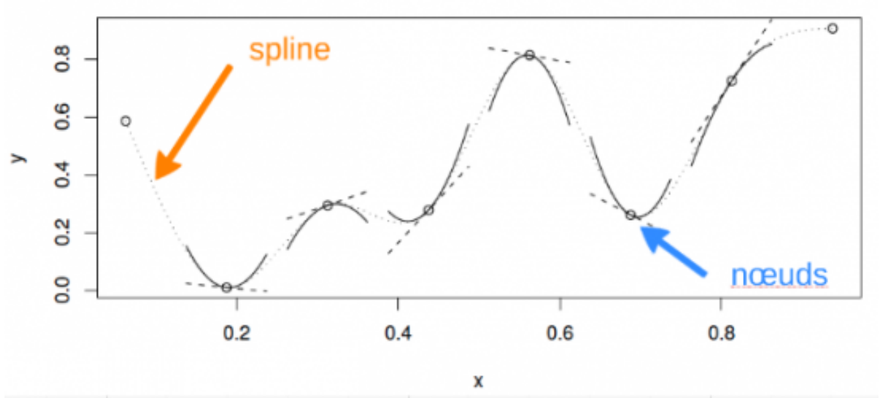


Figure 1. Graphe acyclique dirigé du prior peNMIG. Source : Scheipl, 2011.

β : vecteur de coefficients du modèle.

α : importance du bloc/effet de la variable.

ξ : redistribue α entre les bases de la spline.



Spline : Fonctions polynomiales définies par morceaux

PRIOR

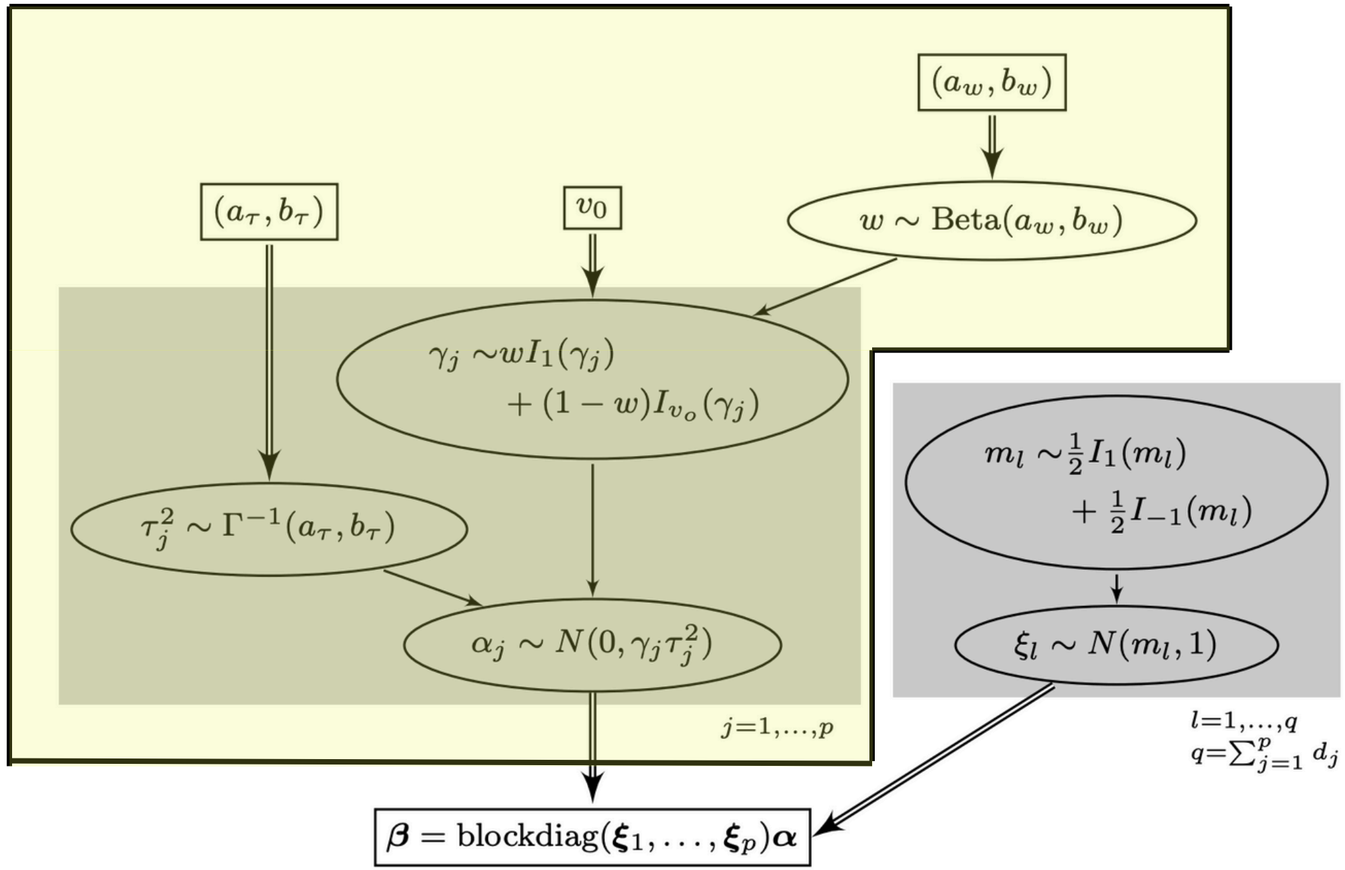
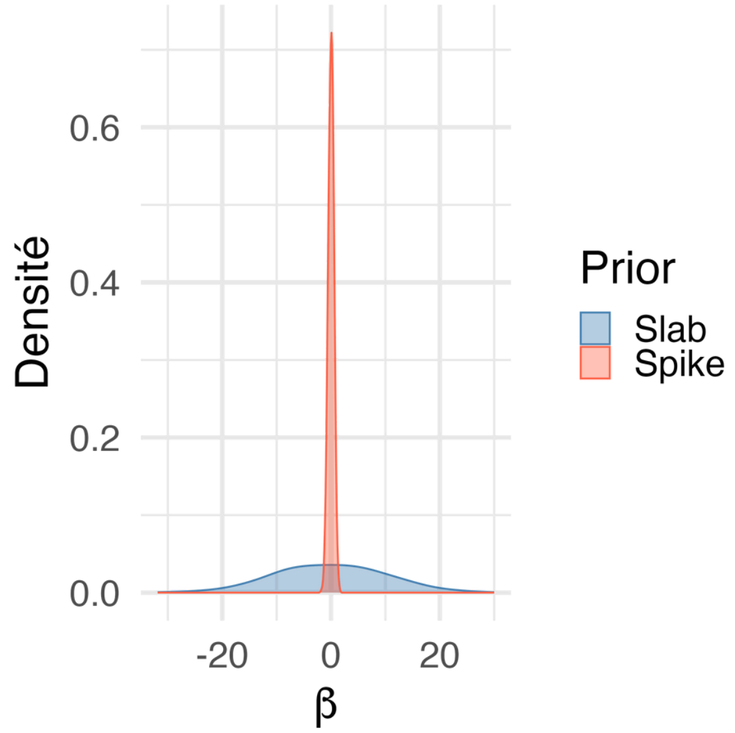


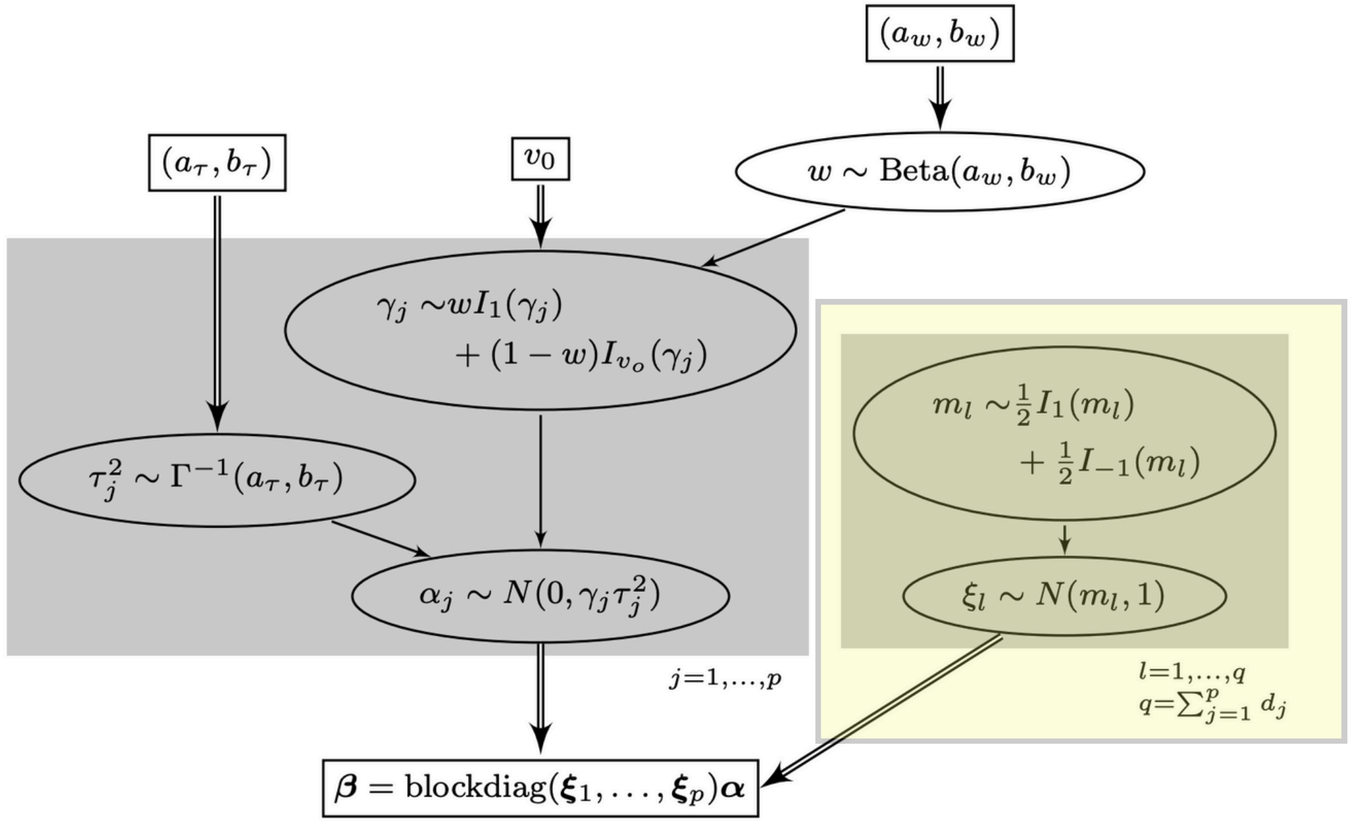
Figure 1. Graphe acyclique dirigé du prior peNMIG. Source : Scheipl, 2011.

α_j : importance du bloc j

- Pour les effets linéaires : effet linéaire de la variable.
- Pour les splines : effet de la spline.



PRIOR



ξ_l : redistribuent α_j entre les différentes composantes du bloc (par ex. les différentes fonctions de base spline).

Figure 1. Graphe acyclique dirigé du prior peNMIG. Source : Scheipl, 2011.

Méthode

Données nimble :
- constante
- data (matrice de design, gauss Legendre)

Modèle nimble:

```
nimbleCode({
#PRIORS
        ...
#LIKELIHOOD
        ...
})
```

Création de l'object R qui représente les noeuds :

nimbleModel()

Décision de la stratégie d'échantillonnage :

configureMCMC()

Construction du MCMC :

buildMCMC()

Compilateur C++ :

compileNimble()

Execution :

runMCMC()

Package R nimble

Méthode

1 FORMULE

Surv(t, status) ~ t + x1 + x2
OU
Surv(t, status) ~ lin(t) + sm(t) + lin(x1) + lin(x2)

2 MATRICE DE DESIGN

Création de matrice de design pour chaque terme du modèle :

- **U() :** intercept, terme non soumis à la sélection

Méthode

1 FORMULE

Surv(t, status) ~ t + x1 + x2
OU
Surv(t, status) ~ lin(t) + sm(t) + lin(x1) + lin(x2)

2 MATRICE DE DESIGN

Création de matrice de design pour chaque terme du modèle :

- **U() :** intercept, terme non soumis à la sélection
- **fct():** terme facteur
  a. création de la matrice de design à partir de la **matrice de contraste**
  b. **Mise à l'échelle** : Norme de Frobenius standardisée = 0.5

Méthode

1 FORMULE

Surv(t, status) ~ t + x1 + x2
OU
Surv(t, status) ~ lin(t) + sm(t) + lin(x1) + lin(x2)

2 MATRICE DE DESIGN

Création de matrice de design pour chaque terme du modèle :

- **U() :** intercept, terme non soumis à la sélection
- **fct():** terme facteur
- **lin():** terme linéaire
  a. **Transformation :** polynome orthogonal de degré 1 (centré)
  b. **Mise à l'échelle :** Norme de Frobenius standardisée = 0.5

Méthode

1 FORMULE

Surv(t, status) ~ t + x1 + x2
OU
Surv(t, status) ~ lin(t) + sm(t) + lin(x1) + lin(x2)

2 MATRICE DE DESIGN

Création de matrice de design pour chaque terme du modèle :

- **U() :** intercept, terme non soumis à la sélection
- **fct():** terme facteur
- **lin():** terme linéaire
- **sm():** terme non-linéaire
  a. **Base Initiale :** B-Splines + Pénalité P (différences finies d'ordre 2)
  b. **Séparation** des composantes constante/linéaire de celles non-linéaires. (projection orthogonale)
  c. **Mise à l'échelle** : Norme de Frobenius standardisée globale = 0.5

Méthode

Données nimble :
- constante
- data (matrice de design, gauss Legendre)

Modèle nimble:

```
nimbleCode({
#PRIORS
                ...
#LIKELIHOOD
                ...
})
```

Création de l'object R qui représente les noeuds :

nimbleModel()

Décision de la stratégie d'échantillonnage :

configureMCMC()

Construction du MCMC :

buildMCMC()

Compilateur C++ :

compileNimble()

Execution :

runMCMC()

Package R nimble

Méthode

Données nimble :
- constante
- data (matrice de design, gauss Legendre)

Modèle nimble:

```
nimbleCode({
#PRIORS
                ...
#LIKELIHOOD
                ...
})
```

Création de l'object R qui représente les noeuds :

nimbleModel()

Décision de la stratégie d'échantillonnage :

configureMCMC()

Construction du MCMC :

buildMCMC()

Compilateur C++ :

compileNimble()

Execution :

runMCMC()

Package R nimble

Méthode

Données nimble :
- constante
- data (matrice de design, gauss Legendre)

Modèle nimble:

```
nimbleCode({
#PRIORS
                ...
#LIKELIHOOD
                ...
})
```

Création de l'object R qui représente les noeuds :

nimbleModel()

Décision de la stratégie d'échantillonnage :

configureMCMC()

Construction du MCMC :

buildMCMC()

Compilateur C++ :

compileNimble()

Execution :

runMCMC()

Package R nimble

Méthode

Paramètres	Sampler	Pourquoi ?	Description
Binaires (gamma, m)	Gibbs sampling pour valeurs binaires	Discret (0 ou 1)	Calcule P(1|...) vs P(0|...) : $p_j = \frac{P(1|\dots)}{P(0|\dots)+P(1|\dots)}$ et tire dans une loi de bernouilli avec une probabilité p_j
Tau	Conjugate (Gibbs)	Prior Gamma + Likelihood Normale	Échantillonnage exact, très rapide.
Autres (alpha, ski, w)	RW (Random Walk)	Conjugaison brisée (Zero Trick / Formules)	Exploration locale "Essai-Erreur" (Metropolis).

Méthode

Données nimble :
- constante
- data (matrice de design, gauss Legendre)

Modèle nimble:

```
nimbleCode({
#PRIORS
                ...
#LIKELIHOOD
                ...
})
```

Création de l'object R qui représente les noeuds :

nimbleModel()

Décision de la stratégie d'échantillonnage :

configureMCMC()

Construction du MCMC :

buildMCMC()

Compilateur C++ :

compileNimble()

Execution :

runMCMC()

Package R nimble

Méthode

Données nimble :
- constante
- data (matrice de design, gauss Legendre)

Modèle nimble:

```
nimbleCode({
#PRIORS
                ...
#LIKELIHOOD
                ...
})
```

Création de l'object R qui représente les noeuds :

nimbleModel()

Décision de la stratégie d'échantillonnage :

configureMCMC()

Construction du MCMC :

buildMCMC()

Compilateur C++ :

compileNimble()

Execution :

runMCMC()

Package R nimble

Méthode

Données nimble :
- constante
- data (matrice de design, gauss Legendre)

Modèle nimble:

```
nimbleCode({
#PRIORS
                ...
#LIKELIHOOD
                ...
})
```

Création de l'object R qui représente les noeuds :

nimbleModel()

Décision de la stratégie d'échantillonnage :

configureMCMC()

Construction du MCMC :

buildMCMC()

Compilateur C++ :

compileNimble()

Execution :

runMCMC()

Package R nimble

ETUDE DE SIMULATION PRELIMINAIRE

Objectif : comparer le modèle bayésien de taux développé (spikeSlabHazard) vs approximation Poisson (spikeSlabGAM).

01 Génération des données et design

- Deux tailles d'échantillon : **500 et 1000 individus**
- **500 jeux** de données indépendants par taille

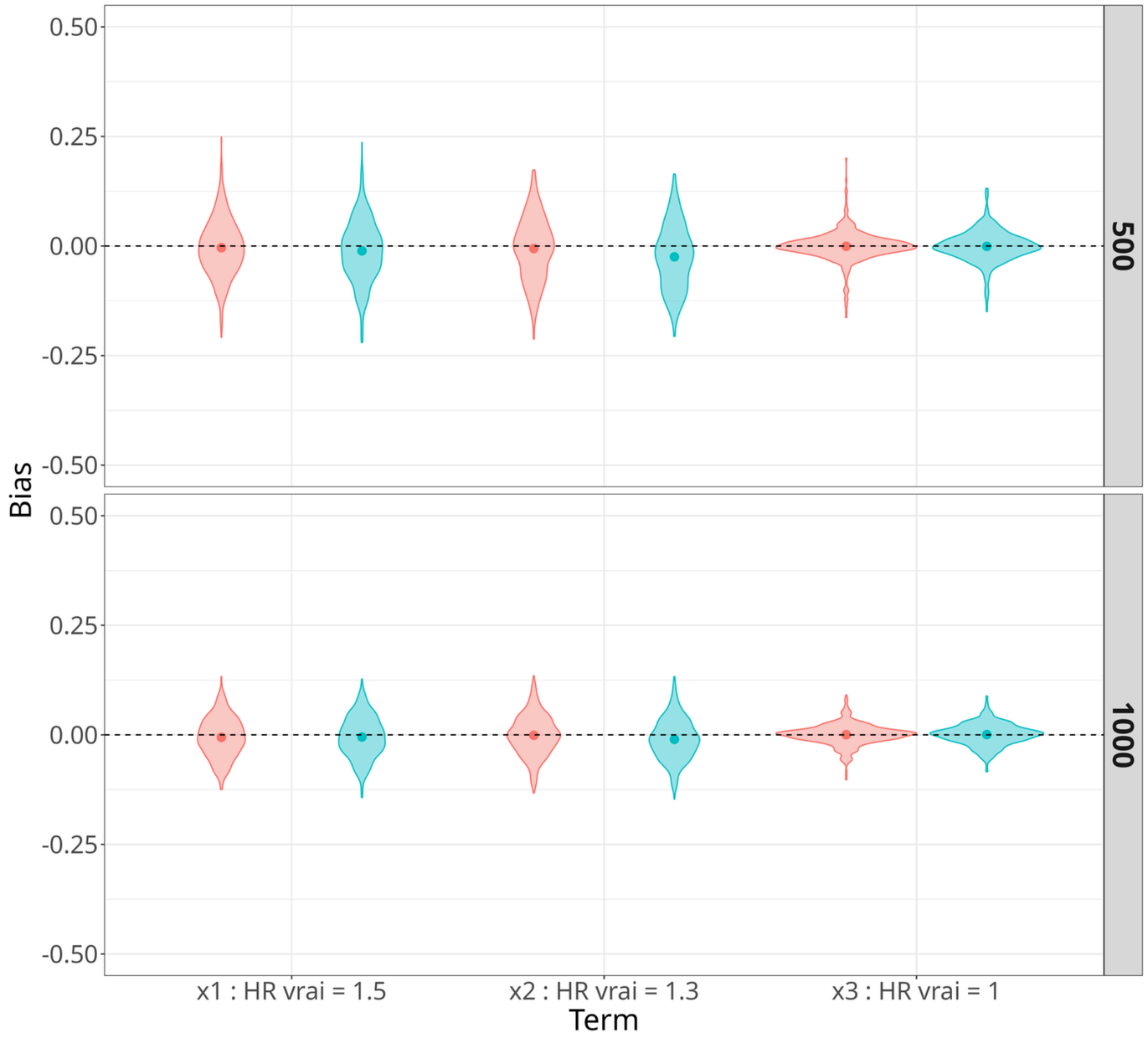
02 Ampleurs d'effets

- **Covariables :** 3 variables ~ N(0,1)
  - Effet réel : **x1 (HR = 1.5)**, **x2 (HR = 1.3)**
  - Pas d'effet : x3
- **Taux de base :** B-splines cubiques (degré 3, nœuds internes à 2.12 et 3.92)

03 Comparaisons

- Critère de précision des estimations :
  - Biais
  - Erreur quadratique moyenne (MSE)
- Probabilité de couverture
- Probabilités d'inclusion

Résultats

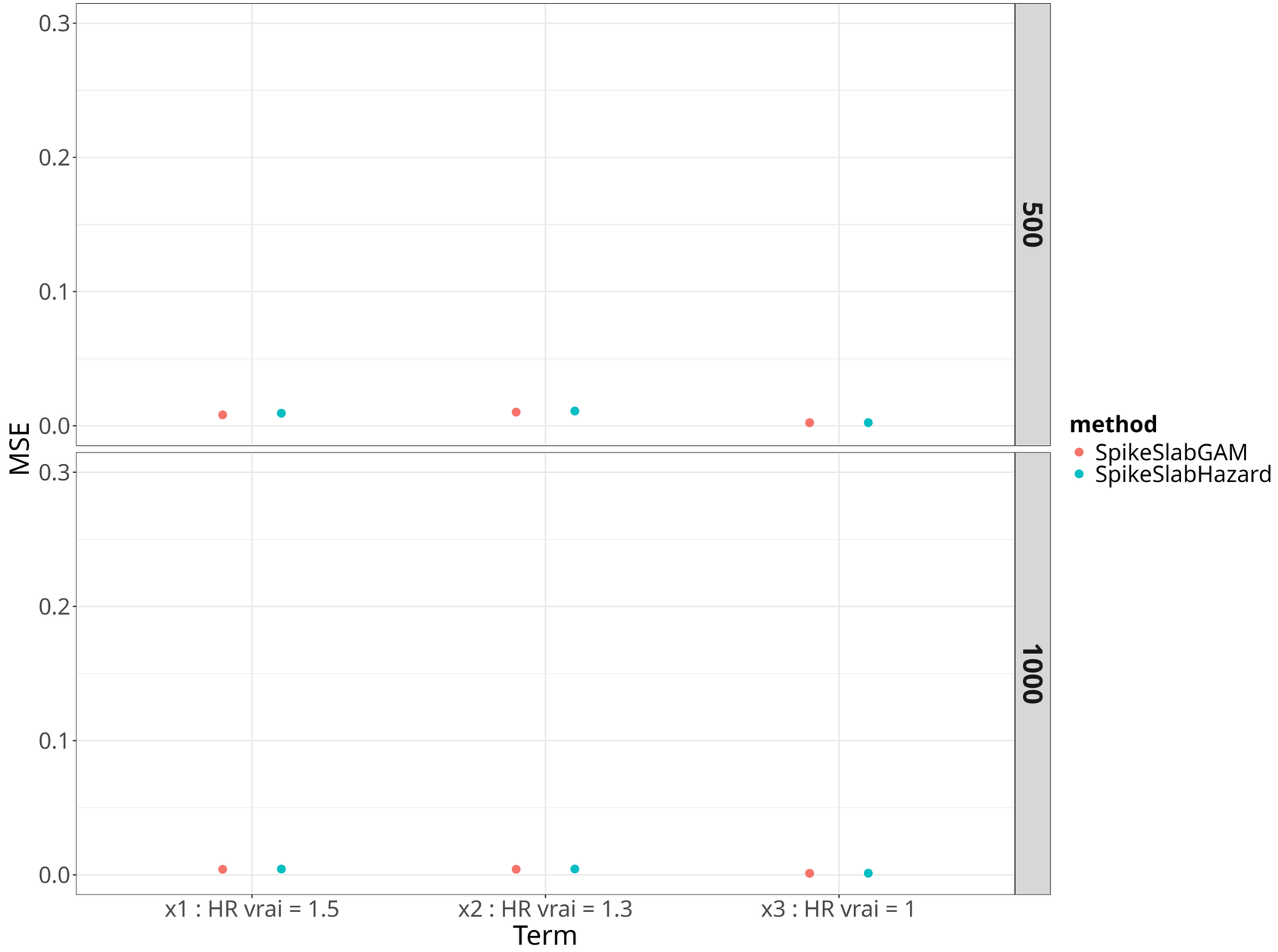


Biais négligeable pour les deux méthodes, biais standardisé<40% :

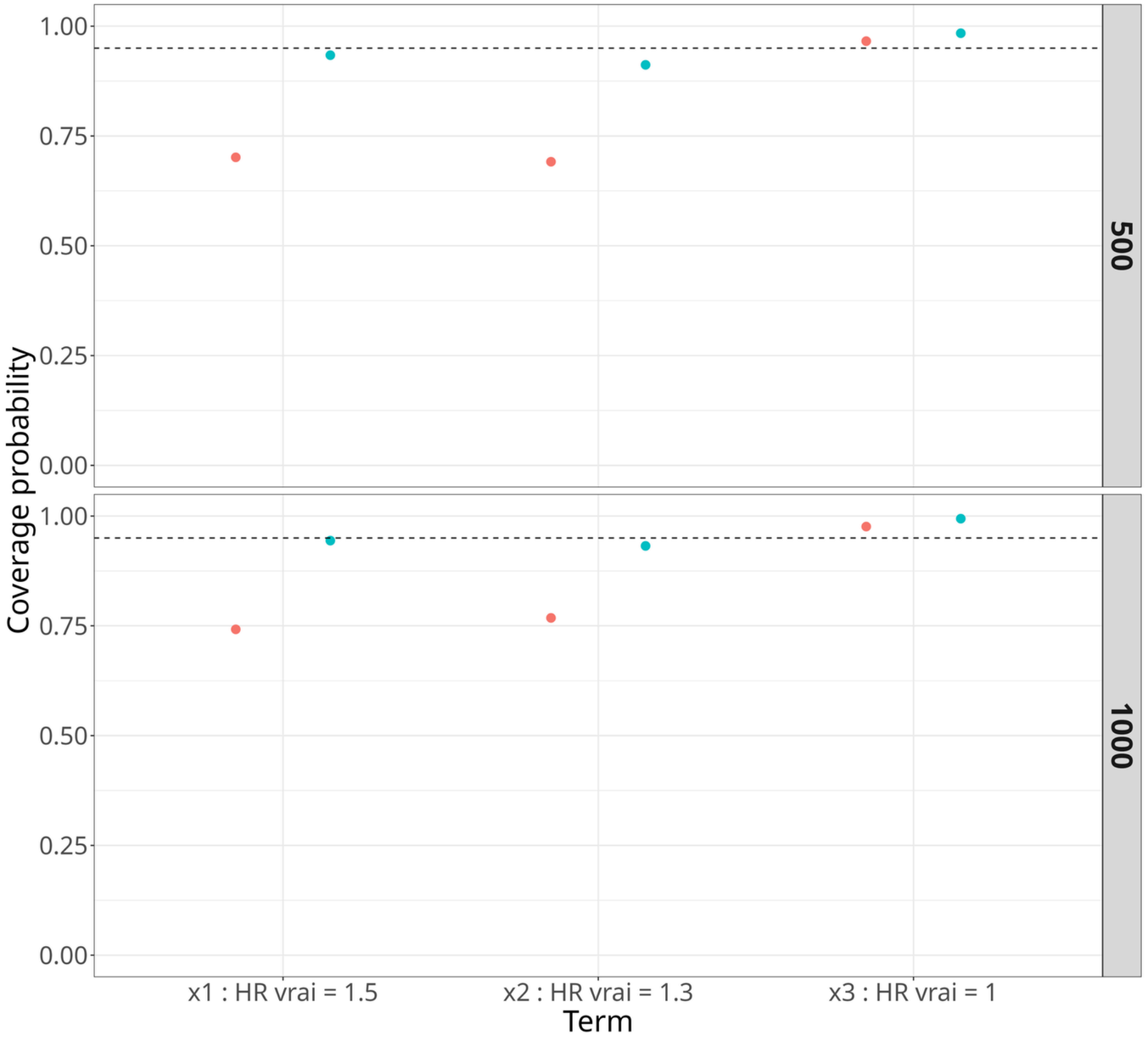
$$\left(\frac{\bar{\hat{\beta}} - \beta}{\mathrm{SE}(\hat{\beta})}\right) * 100$$

method
- SpikeSlabGAM
- SpikeSlabHazard

Résultats


0.3
0.2
0.1
0.0
MSE
500
1000
x1 : HR vrai = 1.5
x2 : HR vrai = 1.3
x3 : HR vrai = 1
Term
method
SpikeSlabGAM
SpikeSlabHazard

Résultats



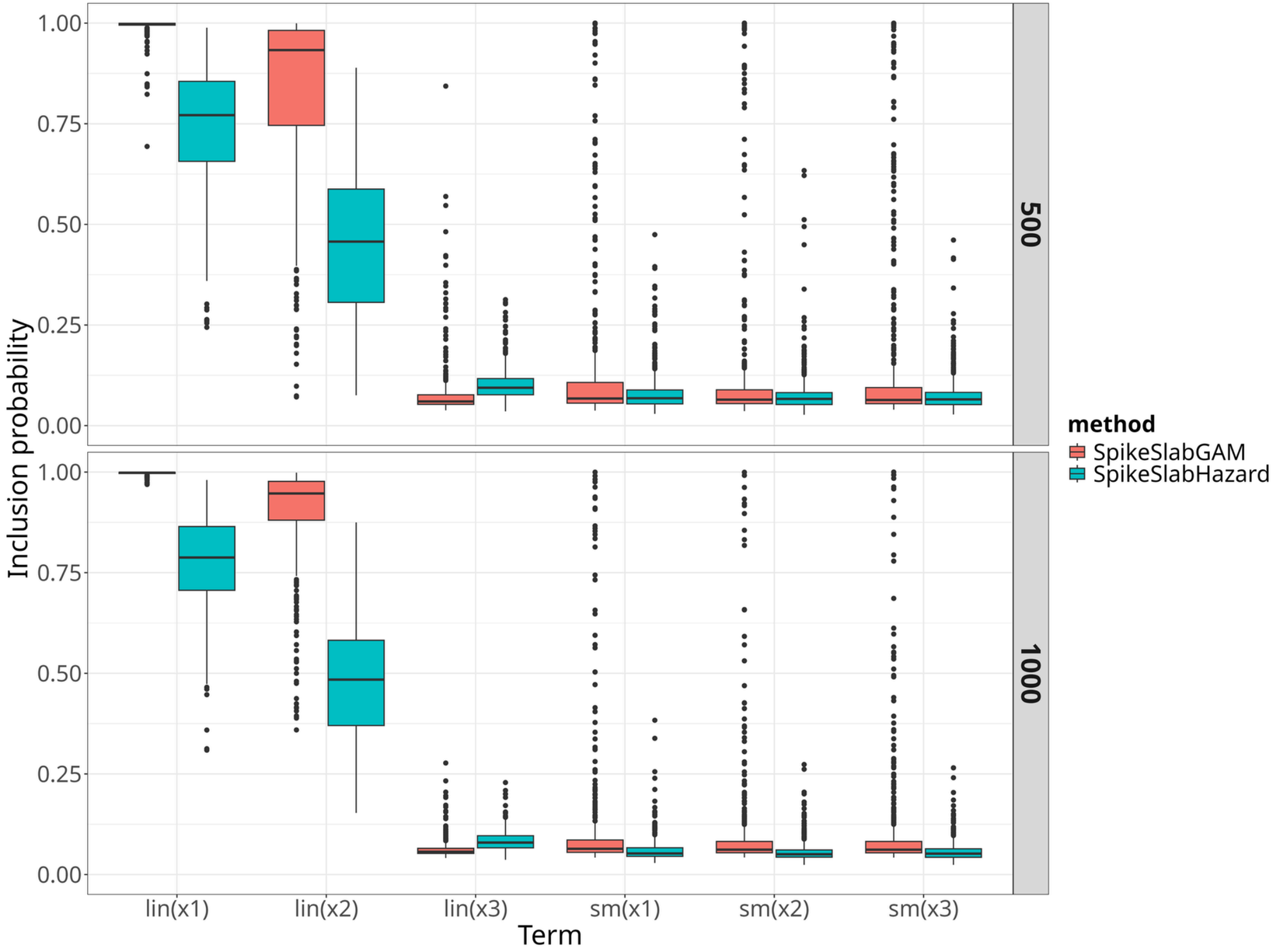
Exemple pour design : n=500

	sd théorique		sd empirique	
x1	0.066	0.032	0.068	0.064
x2	0.065	0.036	0.072	0.071
x3	0.044	0.019	0.0345	0.0341

Différence de probabilité de couverture dû à une différence entre sd théorique et empirique.

method
- SpikeSlabGAM
- SpikeSlabHazard

Résultats

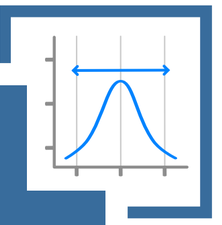

1.00
0.75
0.50
0.25
0.00
Inclusion probability
500
1000
lin(x1)
lin(x2)
lin(x3)
sm(x1)
sm(x2)
sm(x3)
Term
method
SpikeSlabGAM
SpikeSlabHazard

Discussion

Biais, MSE

- ✅ Spike Slab Hazard : négligeable
- ✅ Spike Slab GAM : négligeable

Probabilité de couverture



- ✅ Spike Slab Hazard : correct
- ❌ Spike Slab GAM : sous-couverture

Probabilité d’inclusion

- ⚖️ Spike Slab Hazard : sélection moins marquée
- ✅ Spike Slab GAM : plus élevée pour les covariables actives

- SpikeSlabGAM → favorise davantage l’inclusion
- SpikeSlabHazard → plus prudent, mais inclusion moins tranchée.

Perspectives

Simulations

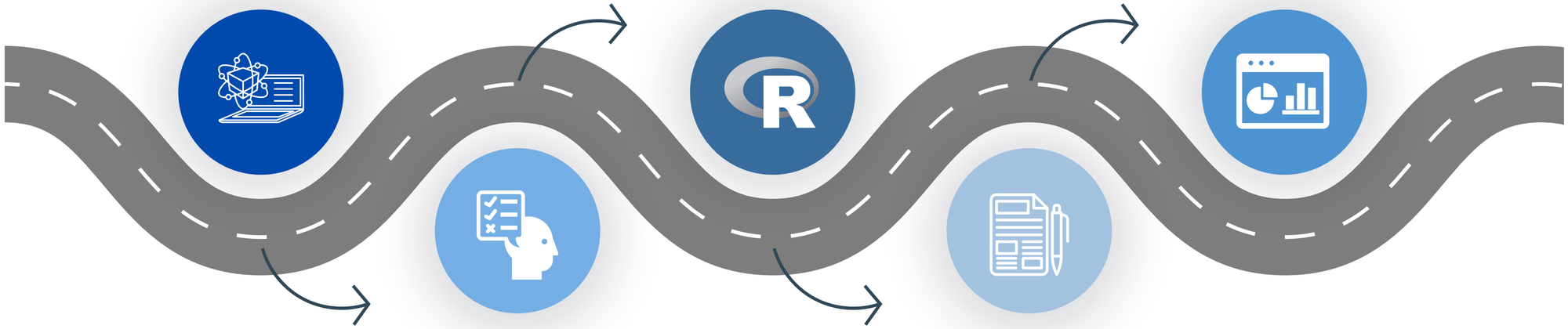
- Evaluer d’autres designs

Package

- Production d’un package R

Application

- Méta-analyse de données d’ECR sur les traitements de l’hypertension artérielle.



Stratégies de seuillage

Publication

- Rédaction d’un article


Merci