Introduction aux événements récurrents en grande dimension

Journées de Biostatistique

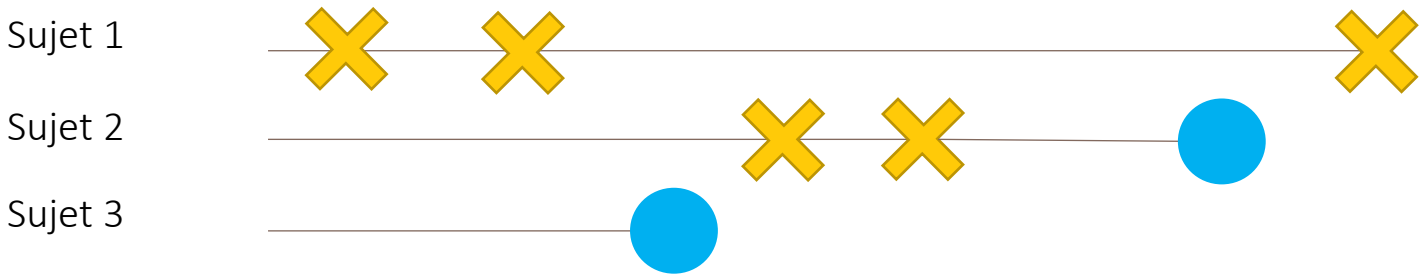
Le 18 novembre 2022

Juliette Murris, doctorante à HeKA (Inria – Inserm),
Sous l'encadrement de Sandrine Katsahian et Audrey Lavenu

Agenda

1. Contexte
2. Objectifs
3. Analyser des événements récurrents
4. Traiter la grande dimension
5. Discussion/Conclusion

Contexte

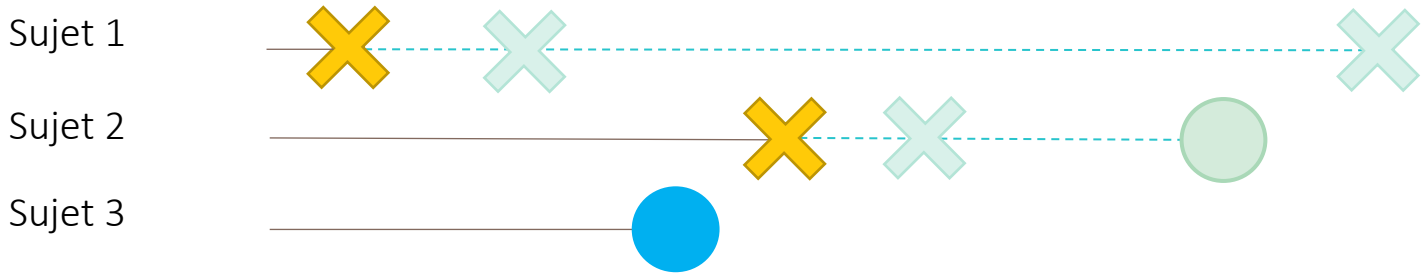


- Plusieurs événements pour chaque individu
- Evolution des facteurs au cours du temps

Evénement récurrent

Censure

Contexte

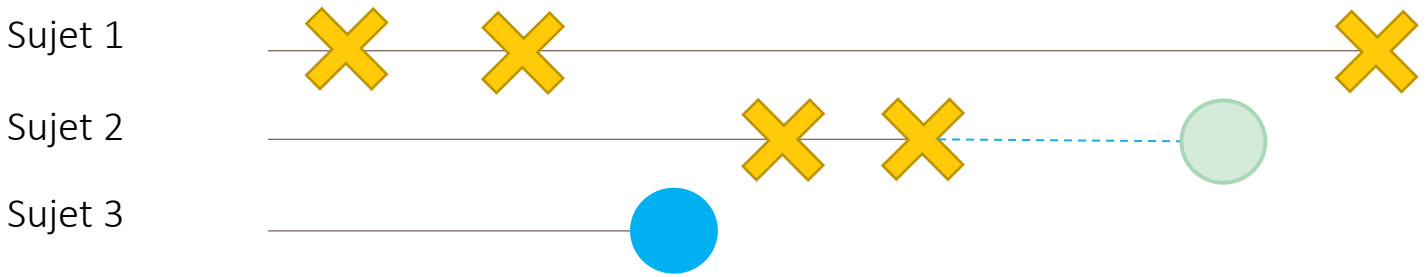


- Approche classique avec le modèle de Cox

Evénement récurrent

Censure

Contexte



- Modèles statistiques pour les événements récurrents
  - Extensions de modèles de Cox : Andersen-Gill, Prentice-William & Petersen, Wei-Lin & Weissfeld
  - Modèles de comptage : processus de Poisson, Binomial négatif

Evénement récurrent

Censure

Contexte

La grande dimension

$p > n$ avec p le nombre de facteurs et n le nombre d'individus

Recours habituel pour traiter ce problème

- Réduction de dimension : ACP, regressions pénalisées (Lasso, Enet, Ridge)
- Machine learning : forêts aléatoires, SVM, reseaux de neurones

Les objectifs aujourd'hui

- Comment analyser les événements récurrents?
- Comment les traiter dans un contexte de grande dimension?

Notations

$\mathbf{X} \in \mathbb{R}^{n*p}$ la matrice des facteurs

β les coefficients associés

$\lambda_0(t)$ la function de risque de base

$Y_i(t)$ une indicatrice pour définir si l'individu i est à risqué au temps t

$T_i = E_i \wedge C_i$ le temps minimal entre l'évenement et la censure

η_i le risqué d'apparition de l'événement

$N_i^*(t)$ le nombre total d'événements sur l'intervalle $[0, t]$

Analyse des événements récurrents

	AG	PWP	WLW	Poisson	NB
Temps jusqu'aux événements	x	x			
Temps entre chaque événement			x		
Nombre total d'événements				x	
Taux d'apparition (pour chaque unité de temps)					x

AG = Andersen-Gill, NB = Négatif binomial, PWP = Prentice, William et Peterson, WLW = Wei-Lin-Weissfeld

Analyse des événements récurrents

	AG	PWP	WLW	Poisson	NB
Temps jusqu'aux événements	x	x			
Temps entre chaque événement			x		
Nombre total d'événements				x	
Taux d'apparition (pour chaque unité de temps)					x

Take home message

La méthodologie pour l'analyse des événements récurrents repose essentiellement sur l'objectif de l'étude / la question scientifique.

AG = Andersen-Gill, NB = Négatif binomial, PWP = Prentice, William et Peterson, WLW = Wei-Lin-Weissfeld

Analyse des événements récurrents

	AG	PWP	WLW	Poisson	NB
Temps jusqu'aux événements	x	x			
Temps entre chaque événement			x		
Nombre total d'événements				x	
Taux d'apparition (pour chaque unité de temps)					x

AG = Andersen-Gill, NB = Négatif binomial, PWP = Prentice, William et Peterson, WLW = Wei-Lin-Weissfeld

Analyse des événements récurrents

	AG	PWP	WLW
Temps jusqu'aux événements	x	x	
Temps entre chaque événement			x

AG – Les événements récurrents au sein des individus sont indépendants et partagent une fonction de risque de base commune

$$\lambda_i(t) = Y_i(t) \times \lambda_0(t) \times \exp(\beta^t X_i)$$

Possibilité de prendre en compte les covariables variant dans le temps et les intervalles de temps à risque discontinus

Mais

L'omission d'une covariable importante pourrait induire une dépendance

AG = Andersen-Gill, PWP = Prentice, William et Peterson, WLW = Wei-Lin-Weissfeld

Analyse des événements récurrents

	AG	PWP	WLW
Temps jusqu'aux événements	x	x	
Temps entre chaque événement			x

PWP = AG stratifié – La strate k considère les $k^{\text{èmes}}$ événements de l'individu i

$$\lambda_{ik}(t) = Y_i(t) \times \lambda_{0k}(t) \times \exp(\beta_k^t X_i)$$

Recommandé lorsque l'on suppose que la survenue du premier événement augmente la probabilité d'une récurrence

Mais

Estimations instables si les risques pour les événements ultérieurs décroissent

AG = Andersen-Gill, PWP = Prentice, William et Peterson, WLW = Wei-Lin-Weissfeld

Analyse des événements récurrents

	AG	PWP	**WLW**
Temps jusqu'aux événements	x	x	
Temps entre chaque événement			x

WLW – Modèle marginal avec une échelle de temps calendaire

$$\lambda_{ik}(t) = Y_i(t) \times \lambda_{0k}(t) \times \exp(\beta_k^t X_i)$$

La dépendance intra-patient pris en compte dans l'estimation de la variance

Mais

Besoin de limiter le nombre d'événements par patient

AG = Andersen-Gill, PWP = Prentice, William et Peterson, WLW = Wei-Lin-Weissfeld

Analyse des événements récurrents

Population : 85 patients atteints d'un cancer de la vessie

Objectif : évaluer l'effet de deux bras de traitement (thiotépa ou placebo) sur la récidive tumorale

HR (95% CI)	AG	PWP	WLW	CPH
Traitement (Réf: placebo)	0.63 (0.40 – 0.99)	0.72 (0.48 – 1.05)	0.56 (0.30 – 1.02)	0.59 (0.32 – 1.10)
Nombre de tumeurs	1.19 (1.05 – 1.35)	1.12 (1.02 – 1.24)	1.23 (1.08 – 1.41)	1.27 (1.09 – 1.47)
Taille de la plus grosse tumeur	0.96 (0.83 – 1.11)	0.99 (0.88 – 1.12)	0.95 (0.79 – 1.14)	1.07 (0.88 – 1.31)

AG = Andersen-Gill, CI = Confidence Interval, CPH = Cox Proportional Hazard, HR = Hazard Ratio, PWP = Prentice, William et Peterson, WLW = Wei-Lin-Weissfeld

Traiter la grande dimension

Dans la littérature

Applications

Très souvent la récurrence est évitée :

Classifieur	Recurrence-free	Temps jusqu'au premier événement

1 seule application avec un réseau de neurones WT-RTT

Méthode issue d'un mémoire de master & non publiée dans un journal à comité de lecture

Traiter la grande dimension

Dans la littérature

Applications

Très souvent la récurrence est évitée :

Classifieur	Recurrence-free	Temps jusqu'au premier événement

1 seule application avec un réseau de neurones WT-RTT

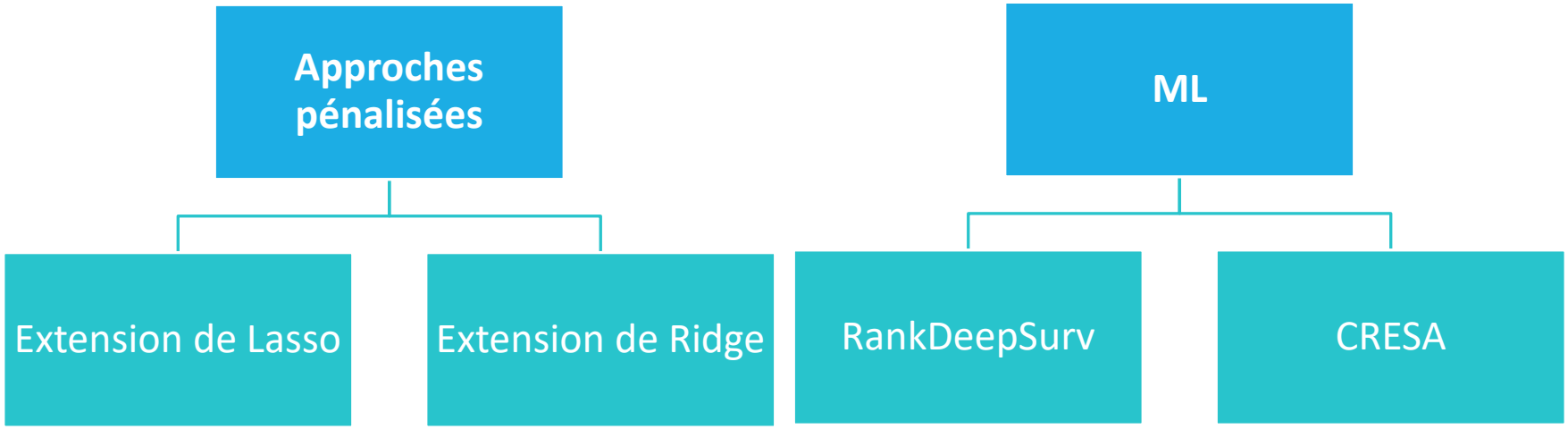
Take home message

En l'absence de recommandations, la revue de la littérature illustre la prudence des auteurs/investigateurs lorsqu'ils traitent des événements récurrents en grande dimension.

Traiter la grande dimension

Dans la littérature

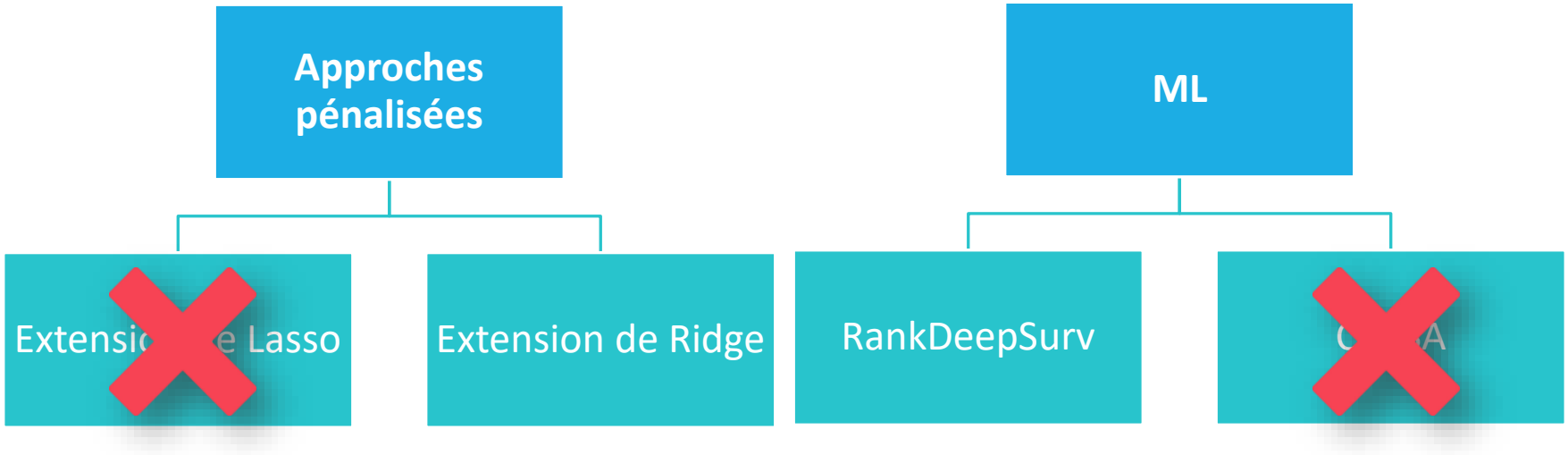
4 méthodes publiées



Traiter la grande dimension

Dans la littérature

4 méthodes publiées



Traiter la grande dimension

Source de données

$n = 100$

$p = 25, 50, 100, 150, 200$

$c = 20\%$ (taux de censure)

$sp = 25\%, 50\%$ (taux de variables actives)

Plan de simulation

Inspiration du package simrec avec introduction du contrôle de la multicolinéarité et de sp

Traiter la grande dimension

Evaluation

C-index de Harrell

$$\hat{\mathbb{C}} = \frac{\sum_{i \neq j} I\{\eta_i < \eta_j\} \times I\{T_i > T_j\} \times \delta_j}{\sum_{i \neq j} I\{T_i > T_j\} \times \delta_j}$$

C-index de Kim

$$\hat{\mathbb{C}}_{rec} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I\{N_i^*(T_i \wedge T_j) > N_j^*(T_i \wedge T_j)\} \times I\{\beta^t X_i > \beta^t X_j\}}{\sum_{i=1}^{n} \sum_{j=1}^{n} I\{N_i^*(T_i \wedge T_j) > N_j^*(T_i \wedge T_j)\}}$$

Traiter la grande dimension

Evaluation

C-index de Harrell

$$\hat{\mathbb{C}} = \frac{\sum_{i \neq j} I\{\eta_i < \eta_j\} \times I\{T_i > T_j\} \times \delta_j}{\sum_{i \neq j} I\{T_i > T_j\} \times \delta_j}$$

C-index de Kim

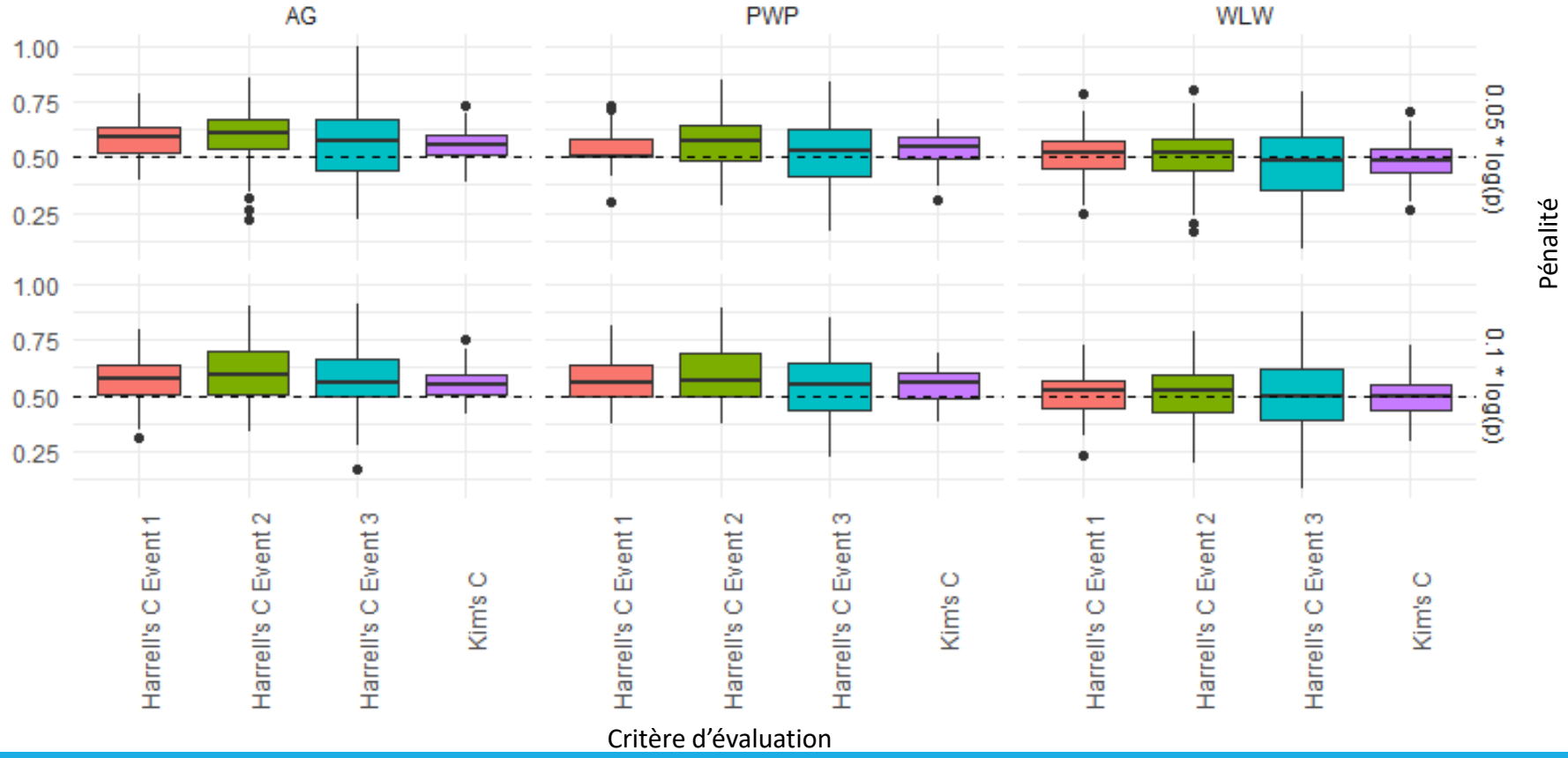
$$\hat{\mathbb{C}}_{rec} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} I\{N_i^*(T_i \wedge T_j) > N_j^*(T_i \wedge T_j)\} \times I\{\beta^t X_i > \beta^t X_j\}}{\sum_{i=1}^{n} \sum_{j=1}^{n} I\{N_i^*(T_i \wedge T_j) > N_j^*(T_i \wedge T_j)\}}$$

Take home message

Il ne semble pas exister de critère unique pour évaluer les méthodes qui traitent les événements récurrents (avec une approche time-to-event)

Traiter la grande dimension

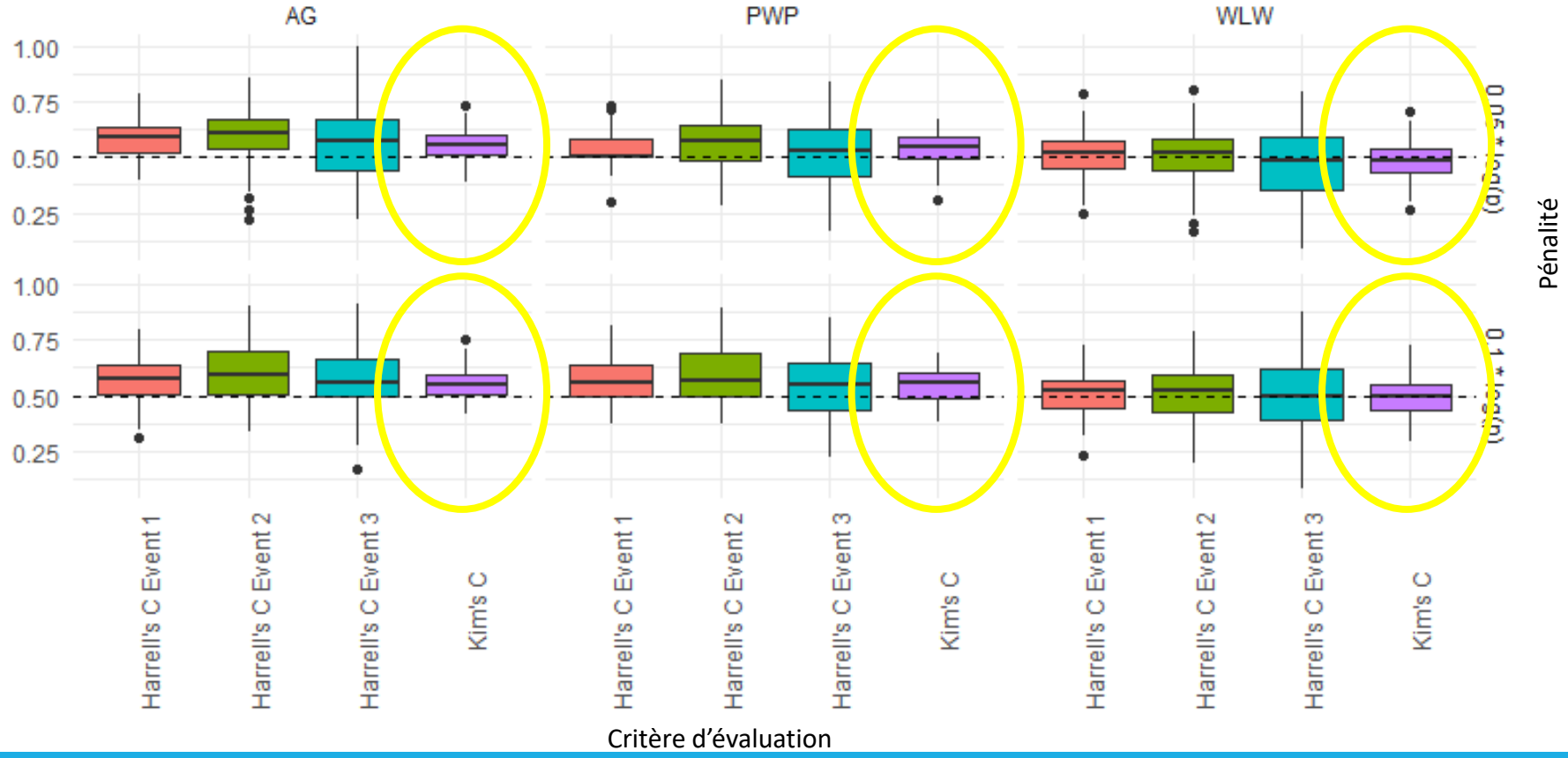
Résultats – $p = 150,\ sp = 25\%$



AG = Andersen-Gill, PWP = Prentice, William et Peterson, WLW = Wei-Lin-Weissfeld

Traiter la grande dimension

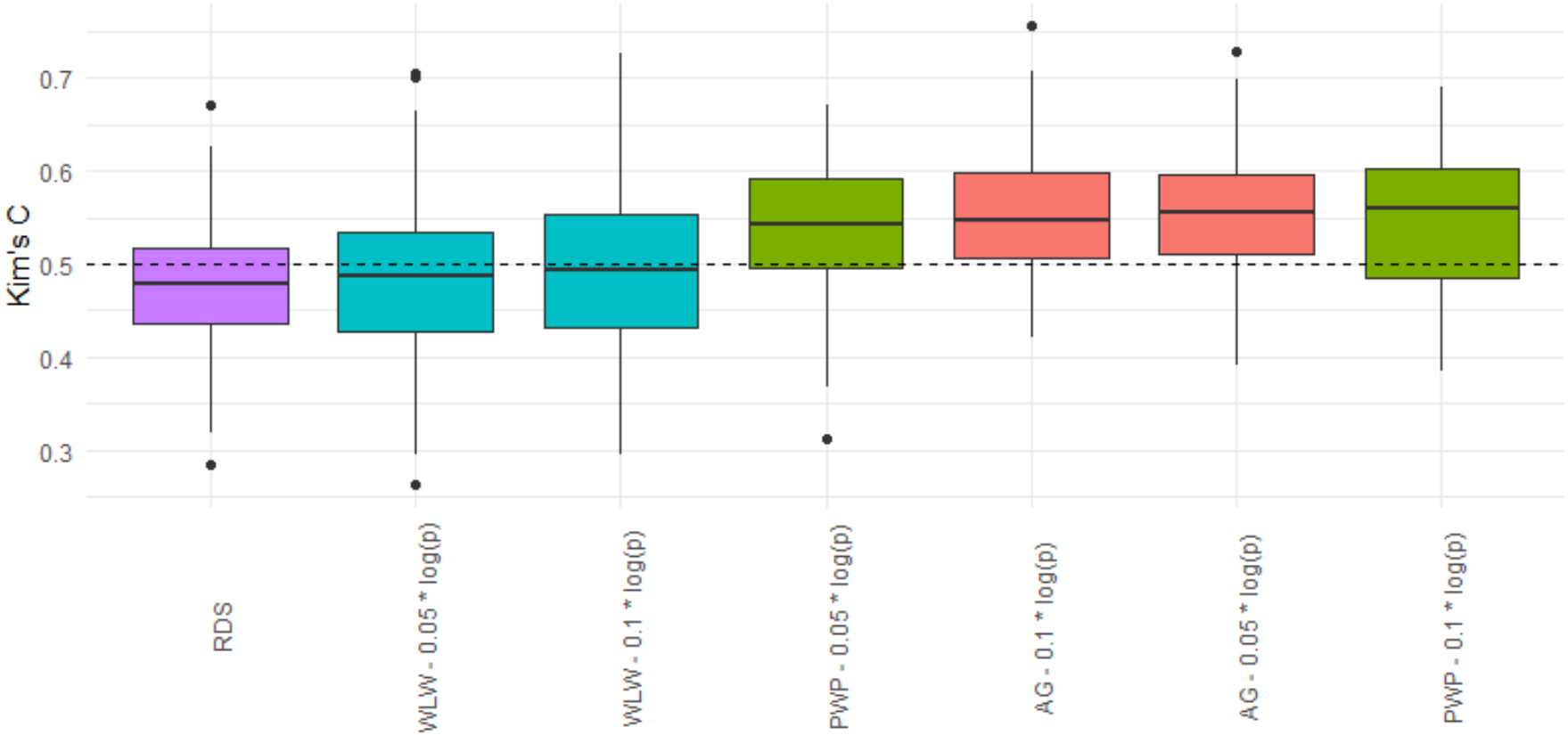
Résultats – $p = 150,\ sp = 25\%$



AG = Andersen-Gill, PWP = Prentice, William et Peterson, WLW = Wei-Lin-Weissfeld

Traiter la grande dimension

Résultats – $p = 150,\ sp = 25\%$



AG = Andersen-Gill, PWP = Prentice, RDS = RankDeepSurv, William et Peterson, WLW = Wei-Lin-Weissfeld

Discussion

Forces

- Identification des dernières approches pour répondre à la problématique des événements récurrents en grande dimension
- 1e confrontation de méthodes standard, d'algorithmes de sélection de variables et d'un réseau de neurones

Limites

- Les hyperparamètres de la méthode de pénalité BAR n'ont pas pu être optimisés
- D'autres mesures d'évaluation pourraient être utilisées

Conclusion

- Aucune méthode ML disponible ne semble plus performante
- Aucune application des méthodes à sélection de variables dans la littérature
- Pas de recommandation en regard de la métrique

Conclusion

- Aucune méthode ML disponible ne semble plus performante
- Aucune application des méthodes à sélection de variables dans la littérature
- Pas de recommandation en regard de la métrique

Final take home message
L'analyse des événements récurrents dans un contexte de grande dimension semble encore à explorer

Références

Amorim LDAF, Cai J. Modelling recurrent events: a tutorial for analysis in epidemiology. Int J Epidemiol. 2015;44(1):324–33.

Andersen PK, Gill RD. Cox's Regression Model for Counting Processes: A Large Sample Study. Ann Stat. 1982;10(4):1100–20.

Cox DR. Regression Models and Life-Tables. J R Stat Soc Ser B Methodol. 1972;34(2):187–202.

Gupta G, Sunder V, Prasad R, Shroff G. CRESA: A Deep Learning Approach to Competing Risks, Recurrent Event Survival Analysis. In: Advances in Knowledge Discovery and Data Mining. Cham: Springer International Publishing; 2019. p. 108–22.

Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. JAMA. 1982;247(18):2543–6.

Kim S, Schaubel DE, McCullough KP. A C-index for recurrent event data: Application to hospitalizations among dialysis patients. Biometrics. 2018;74(2):734–43.

Jing B, Zhang T, Wang Z, Jin Y, Liu K, Qiu W, et al. A deep survival analysis method based on ranking. Artif Intell Med. 2019;98:1–9.

Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. Biometrika. 1981;68(2):373–9.

Wei LJ, Lin DY, Weissfeld L. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. J Am Stat Assoc. 1989;84(408):1065–73.

Wu TT. Lasso penalized semiparametric regression on high-dimensional recurrent event data via coordinate descent. J Stat Comput Simul. 2013;83(6):1145–55.

Zhao H, Sun D, Li G, Sun J. Variable selection for recurrent event data with broken adaptive ridge regression. Can J Stat. 2018;46(3):416–28.