Journal de la Société Française de Statistique
Vol. 151 No. 1 (2010)

Choix d'estimateurs basé sur le risque de Kullback-Leibler

Benoit Liquet[1]

Title: Choice of estimators based on Kullback-Leibler risk

Résumé : Le choix d'estimateurs est un point crucial en statistique. Le critère le plus connu dans ce domaine est le critère proposé par Akaike. Il a été présenté comme une estimation, à une constante près, du risque de Kullback-Leibler. Cependant une valeur précise du critère d'Akaike n'a pas d'interprétation directe et la variabilité de ce critère est souvent ignorée. Nous exposons plusieurs approches pour estimer des différences de risques de Kullback-Leibler. Les critères proposés peuvent être utilisés dans un contexte paramétrique, non-paramétrique ou semi-paramétrique. Une extension de ces critères aux cas de données incomplètes est présentée. Plusieurs applications dans le cadre de donnée de survie sont décrits : choix d'estimateurs lisses pour la fonction de risque, choix entre estimateurs issus du modèle à risques proportionnels et de modèle stratifié, et choix entre estimateurs issus de modèle markovien et non-markovien. Dans le prolongement de ces travaux, des critères sont définis pour le choix d'estimateurs basés sur des observations différentes.

Abstract: Estimators choice is a crucial topic in statistics. The most famous criterion is the Akaike information criterion. It has been constructed as an approximation, up to a constant, of the Kullback-Leibler risk. However, a precise value of the Akaike criterion has no direct interpretation and its variability is often ignored. We propose several approaches to estimate Kullback-Leibler risks. The criteria defined can be used in a parametric, non-parametric or semi-parametric context. An extension of these criteria for incomplete data is presented. The issue of the choice of estimators in the presence of incomplete data is described. Several applications in the survival framework is described: smooth estimators choice for the hazard function, estimators choice from proportional hazard model and stratified model, and estimators choice for markov model and non markov model. Finally, several criteria are defined for selecting estimators based on different observations.

Mots-clés : AIC, données censurées, risque de Kullback-Leibler, sélection de modèle, validation croisée
Keywords: AIC criterion, incomplete data, Kullback-Leibler risk, model selection, cross-validation
Classification AMS 2000 : 62B10, 62F07, 62N99, 62P10

[1] ISPED, INSERM U 897
Université Victor-Ségalen Bordeaux 2
33076 Bordeaux cedex. E-mail : benoit.liquet@isped.u-bordeaux2.fr

 ISSN: 2102-6238

1. Introduction

La sélection d'estimateurs est un point crucial en statistique. Le critère principal dans ce domaine est le critère AIC proposé par Akaike [2]. L'AIC ("An Information Criterion") est un critère d'information dans le sens qu'il a été construit comme une estimation, à une constante près, de l'information de Kullback-Leibler [20]. Ce critère a eu un impact majeur dans l'application des méthodes statistiques, voir la présentation de DeLeuwe [13]. De nombreuses variantes du critère ont été proposées. Nous pouvons citer, en particulier l'EIC ([18] ; [31]), qui fait usage du bootstrap pour estimer le risque de Kullback-Leibler. Ce critère est étendu au choix d'estimateurs semi-paramétriques par Liquet et al. [25]. La validation croisée basée sur la vraisemblance (LCV) a également été proposée comme une approximation du risque de Kullback-Leibler, avec l'avantage que ce critère peut être utilisé dans un contexte paramétrique ou semi-paramétrique ([10] ; [23]). Ce critère permet en particulier de choisir le paramètre de lissage pour des estimateurs lisses tel que les estimateurs à noyaux ou les estimateurs de vraisemblance pénalisée. Dans le contexte d'estimateurs lisses (voir Hastie et Tibshirani [15] ; Silverman [32]), le paramètre de lissage peut aussi être choisi par validation croisée (CV) ou une version approchée (GCV) (voir Craven and Wahba [12]). D'autres alternatives au AIC ont été proposées comme le BIC ([30]) ou des approches fondées sur la notion de complexité (Bozdogan [5]). Pour plus de détails sur les alternatives à l'AIC nous nous référons aux monographies de Linhart et Zucchini [22], Burnham et Anderson [7], Konishi et Kitagawa [19] et Claeskens et Hjort [8].

Par ailleurs, il est connu que les critères du type AIC ne sont pas performants lorsque le nombre de modèles est trop grand (par exemple : la sélection de variable en régression gaussienne [4]). En effet, la conception de ces critères repose sur une approximation asymptotique qui suppose que la liste des modèles est fixée tandis que n tend vers l'infini. Il se trouve que dans des problèmes tels que la sélection de variable ou la detection de rupture, il est souhaitable de laisser croître la taille des modèles avec le nombre d'observations. Depuis une dizaine d'année, une nouvelle approche dite "non asymptotique" a été initiée dans [3] et [4] et pour résoudre de tel type de problèmes. L'idée est de pénaliser un critère de type log-vraisemblance de façon à garantir une performance de sélection optimale. Pour une description pédagogique de cette approche, nous nous référons à l'article de Massart ([27]). Ce type d'approche a entre autres été utilisé dans le cadre de données censurées à droite (voir Brunel et Comte [6]) pour le choix de la dimension des estimateurs par projection de la fonction de hasard.

Dans ce papier, nous considérons l'asymptotique où la liste des modèles est fixée tandis que n tend vers l'infini et nous nous focalisons sur le choix d'estimateurs et non sur le choix de modèles. Cette distinction entre choix de modèles et choix d'estimateurs n'est pas toujours très claire et il est souvent fait un abus de langage dans les applications. Par exemple, le critère AIC est défini pour des estimateurs du maximum de vraisemblance. Ainsi, pour deux modèles différents correspondent deux estimateurs du maximum de vraisemblance différents et choisir l'estimateur au sens du AIC revient également à faire un choix entre les deux modèles considérés. Cependant, à un modèle peut être associé plusieurs procédures d'estimations conduisant à plusieurs estimateurs. Le choix se fera en déterminant le risque associé à chaque estimateur. De plus, nous verrons en section 2.3, qu'un estimateur provenant d'un modèle peut être préféré (au sens d'un risque ultérieurement défini) à un estimateur provenant d'un autre modèle pour une quantité d'information alors que pour une quantité d'information différente l'estimateur du second modèle sera préféré.

 ISSN: 2102-6238

Principe de la sélection d'estimateurs

A partir d'un élément aléatoire $\mathscr{W}$ (vecteur d'observations) de loi inconnue P^*, un des objectifs du statisticien est d'approcher P^*. Le principe est alors de proposer un modèle $\mathscr{M}^1$ défini par un ensemble de lois possibles $\mathscr{M}^1 = \{P^\theta, \theta \in \Theta\}$. Plusieurs approches sont envisageables : l'approche paramétrique où Θ est un espace euclidien, l'approche non-paramétrique où Θ est un espace fonctionnel et l'approche semi-paramétrique où $\Theta = \Theta_1 \times \Theta_2$ avec Θ_1 espace euclidien et Θ_2 espace fonctionnel. La démarche classique en statistique est alors de supposer qu'il existe $\theta^* \in \Theta$ tel que $P^* = P^{\theta^*}$ et à partir de $\mathscr{W}$ de construire P^* sous la forme $P^{\hat{\theta}(\mathscr{W})}$. Comme P^* est inconnue, il est difficile de vérifier cette hypothèse ; de plus dans les applications on peut penser qu'elle n'est pas vérifiée. Il en découle une erreur, appelée "erreur d'approximation" et quantifiée par un risque de "mauvaise-spécification". Ce risque est souvent plus important pour un modèle "petit". D'un autre côté, plus le modèle est riche et plus on a de chance de disposer d'une loi P^{θ_0} proche de la réalité, mais en contre-partie $P^{\hat{\theta}(\mathscr{W})}$ a alors un risque d'être "loin" de P^{θ_0} et donc de P^*. Ce risque est appelé "risque d'estimation" ou "risque statistique". Nous schématisons cette situation pour deux modèles $\mathscr{M}^1$ et $\mathscr{M}^2$ par la figure 1 ci-après.

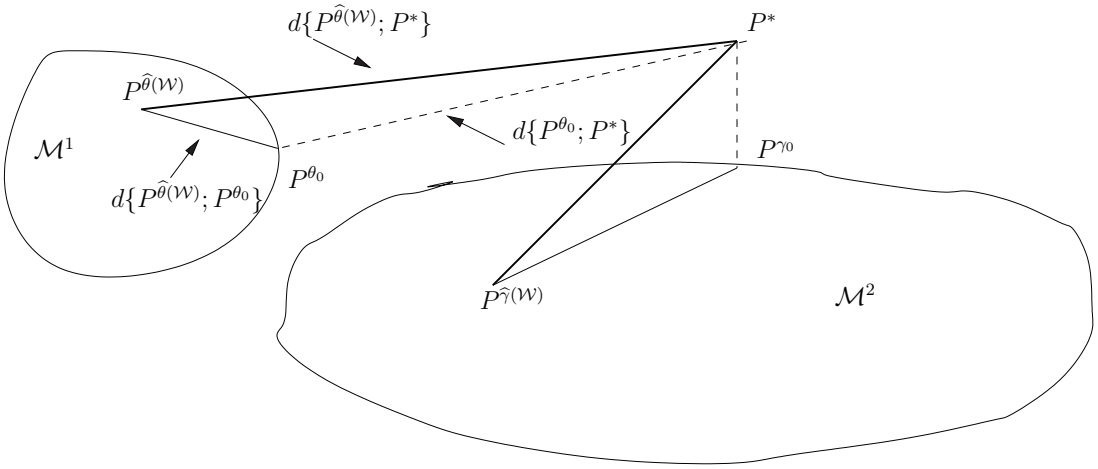

FIGURE 1. *Représentation des risques de mauvaise-spécification et des risques statistiques dans le choix d'estimateurs issus de deux modèles $\mathscr{M}^1$ et $\mathscr{M}^2$*

A la figure 1, les différentes notions de "risque de mauvaise-spécification" et "risque statistique" sont symbolisées par les termes $d\{\cdot;\cdot\}$ qui sera plus précisément défini par la suite comme l'espérance d'une fonction de perte. Par exemple, pour le modèle $\mathscr{M}^1$, la quantité $d\{P^{\theta_0};P^*\}$ représente le "risque de mauvaise-spécification" et s'annule lorsque la vraie loi appartient au modèle considéré. Ce terme traduit la qualité d'approximation du modèle. La quantité $d\{P^{\widehat{\theta}(\mathscr{W})};P^{\theta_0}\}$

 ISSN: 2102-6238

est le "risque statistique". Ce terme est dû à l'estimation faite et est d'autant plus grand que le modèle considéré est grand. Il traduit l'erreur d'estimation. L'art de la sélection d'estimateurs est de définir un estimateur associé à un modèle qui réalise le meilleur compromis entre le "risque de mauvaise-spécification" et "le risque statistique". Nous avons également, schématisé à la figure 1, les termes de "risque statistique" et de "risque de mauvaise-spécification" dans le modèle $\mathscr{M}^2 = \{P^\gamma, \gamma \in \Gamma\}$. Un moyen de choisir entre les deux estimateurs des deux modèles est de déterminer les 2 mesures $d\{P^{\hat{\theta}(\mathscr{W})}; P^*\}$ et $d\{P_k^{\hat{\gamma}(\mathscr{W})}; P^*\}$ représentant l' "éloignement" des deux estimateurs $P^{\hat{\theta}(\mathscr{W})}$ et $P^{\hat{\gamma}(\mathscr{W})}$ à la réalité P^*. L'objectif est donc de déterminer le meilleur estimateur au sens d'une "mesure de proximité" défini par un risque. Plusieurs mesures comme la distance de Hellinger, la distance en variation totale ([14], [34]) ou encore le risque de Kullback-Leibler sont envisageables pour quantifier $d\{P^{\widehat{\theta}(\mathscr{W})}, P^*\}$.

Nous proposons, dans ce papier, de promouvoir les critères de sélection basés sur le risque de Kullback-Leibler. Les critères pratiques de sélection, découlant du risque de Kullback-Leibler, permettent de résoudre de nombreuses problématiques concrètes notamment dans le domaine de la biostatistique.

Exemple d'application : modélisation du risque de démence

Nous reprenons ici l'application traités par Liquet et al. [26] sur la modélisation du risque de démence des personnes agées. Les données considérées proviennent de l'étude PAQUID [21] (QUID sur les Personnes Agées) dont l'objectif général est d'étudier le vieillissement cérébral et fonctionnel après 65 ans. Les données concernent 3675 sujets non déments à l'entrée dans la cohorte vivant à domicile dans deux départements du sud-ouest de la France (Gironde et Dordogne). Les sujets sont vus six fois ou moins, entre 1988 et 2000 ; 431 cas incidents de démence ont été observées au cours du suivi. Le risque de développement d'une démence est modélisé en fonction de l'âge. Deux variables explicatives ont été considérés : le sexe et le niveau d'études. L'échantillon était constitué de 2.133 femmes et 1.542 hommes. Le niveau d'éducation était réparti en deux catégories : avoir le certificat d'étude/ne pas avoir le certificat d'étude. Plusieurs estimateurs sont alors envisageables pour représenter le risque de démence en tenant compte des deux variables explicatives : estimateurs issus de modèles à risques proportionnels ou estimateurs issus de modèles stratifiés. Le critère de sélection basé sur le risque de Kullback-Leibler peut permettre de choisir entre les différents estimateurs considérés. Dans le but d'avoir une estimation lisse de la fonction de risque, la méthode d'estimations choisie est la vraisemblance pénalisée [17]. Cette méthode d'estimation dépend du calibrage d'un paramètre de lissage qui sera aussi sélectionné par le critère de sélection. Cet exemple d'application sera traité dans la section 4.2.

Dans la suite du papier, nous présentons à la section 2 une théorie générale autour du risque de Kullback-Leibler. Cette présentation est faite en terme de densité de probabilité de variable aléatoire. La section 3 est consacrée à la présentation des critères pratiques de sélection dans des situations standards paramétriques et/ou non-paramétriques. Une extension de ces critères aux cas de données incomplètes est présentée en section 4. Plusieurs applications dans le cadre de donnée de survie sont décrits : choix d'estimateurs lisses pour la fonction de risque, choix entre estimateurs issus du modèle à risques proportionnels et de modèle stratifié, et choix entre estimateurs issus de modèle markovien et non-markovien. La section 5 présente des critères de

 ISSN: 2102-6238

sélection dans un cas non-standard où les estimateurs ne sont pas définis à partir des mêmes informations. Dans ce contexte, deux exemples d'applications biomédicales sont présentés. Nous concluons en section 6.

2. Théorie générale du risque de kullback-Leibler

2.1. Le risque de Kullback-Leibler

Le risque de Kullback-Leibler ([20]) permet de mesurer l'écart entre deux densités de probabilité f^* et g :

$$I(g;f^*) = \int f^*(x) \log \left\{ \frac{f^*(x)}{g(x)} \right\} dx.$$

Cette mesure n'est pas une distance car elle ne vérifie pas l'inégalité triangulaire et n'est pas symétrique. En revanche, le risque de Kullback-Leibler est positive ou nul, $I(g;f^*) \geq 0$, avec égalité si et seulement si $f^* = g$ presque partout. Ainsi, cette quantité permet de quantifier l'information perdue quand g est utilisée pour approcher la réalité inconnue représentée par la densité de probabilité f^*.

Considérons maintenant, un modèle paramétrique $(g^\theta)_{\theta\in\Theta}, \Theta \subset \Re^p$. Le modèle (g^θ) est dit bien spécifié si $f^* \in (g^\theta)$ et mal-spécifié dans le cas contraire. Il est alors possible de quantifier le risque de la densité g^θ à f^* par $I(g^\theta;f^*)$. Ainsi, nous pouvons choisir la valeur de θ qui minimise ce risque. Nous supposons qu'il existe une valeur θ_0 qui minimise $I(g^\theta;f^*)$. Le risque de mauvaise spécification du modèle (g^θ) est défini par :

$$I(g^{\theta_0};f^*) = \min_{\theta\in\Theta} I(g^\theta;f^*).$$

Il est évident que si le modèle est bien spécifié $I(g^\theta;f^*)$ atteint son minimum en θ^* et le risque de mauvaise spécification est nul.

Considérons un échantillon d'observations i.i.d. $\mathscr{W}_n = (W_1,\ldots,W_n)$ où $W_i \sim f^*$. Puisque f^* est inconnue, nous cherchons la densité de probabilité $g^\theta \in (g^\theta)$ la plus proche de f^* au sens du risque de Kullback-Leibler, c'est à dire une densité de probabilité qui minimise $I(g^\theta, f^*)$. Nous avons déjà noté cette densité de probabilité g^{θ_0} mais nous ne pouvons pas pratiquement la trouver car f^* est inconnue. Chercher la densité g^θ qui minimise $I(g^\theta;f^*)$ est équivalent à chercher la densité qui rend maximum $\int f^*(x) \log g^\theta(x) dx = \mathrm{E}_{f^*}\left[\log g^\theta(W)\right]$ où la variable W est considérée comme une nouvelle observation indépendante des W_i et de loi f^*. A partir de la loi des grands nombres on obtient que :

$$n^{-1} \sum_{i=1}^{n} \log g^\theta(W_i) \to \mathrm{E}_{f^*}\left[\log g^\theta(W)\right].$$

La log-vraisemblance divisée par n (terme de gauche) est un estimateur naturel du terme à maximiser. Ainsi, l'estimateur du maximum de vraisemblance, noté ici $g^{\hat{\theta}}$, minimise l'estimateur naturel du risque de Kullback-Leibler.

 ISSN: 2102-6238

Nous pouvons évaluer la performance de cet estimateur en utilisant une extension du risque de Kullback-leibler :

$$\mathrm{EKL}(g^{\hat{\theta}}; f^*) = \mathrm{E}_{f^*}\left[\log \frac{f^*(W)}{g^{\hat{\theta}}(W)}\right].$$

La différence avec le risque de Kullback-Leibler classique est que $g^{\hat{\theta}}$ est aléatoire ; ainsi, dans ce cas, $\mathrm{EKL}(g^{\hat{\theta}}, f^*)$ est l'espérance de la divergence de Kullback-Leibler entre $g^{\hat{\theta}}$ et f^*. La définition de ce critère n'étant pas spécifique à un estimateur du maximum de vraisemblance, il peut être utilisé pour l'évaluation de la performance de n'importe quel estimateur (noté par la suite $g^{\hat{\theta}}$).

2.2. Décomposition du risque de Kullback-Leibler

Le risque $\mathrm{EKL}(g^{\hat{\theta}}; f^*)$ est plus grand que le risque de mauvaise spécification. En effet ce critère se décompose comme suit :

$$\mathrm{EKL}(g^{\hat{\theta}}; f^*) = I(g^{\theta_0}, f^*) + \mathrm{E}_{f^*}\left[\log \frac{g^{\theta_0}(W)}{g^{\hat{\theta}}(W)}\right].$$

Le premier terme de la décomposition correspond au *risque de mauvaise spécification* ($= I(g^{\theta_0}; f^*)$) et le second terme est appelé *risque statistique*. Ainsi le risque total est la somme du risque de mauvaise spécification et du risque statistique. Le risque statistique peut être, dans certain cas, bien estimé alors que le risque de mauvaise spécification n'est pas estimable.

Exemple 1. *Dans des modèles paramétriques ($\Theta \subset \mathbb{R}^p$) et pour des estimateurs issus du maximum de vraisemblance, [22] et [11] montrent que*

$$\mathrm{EKL}(g^{\hat{\theta}}; f^*) = \mathrm{E}_{f^*}\left[\log \frac{f^*(W)}{g^{\theta_0}(W)}\right] + \frac{1}{2} n^{-1} Trace(I^{-1}J) + o(n^{-1}), \tag{1}$$

où I est la matrice d'information et J est la variance des scores, toutes les deux calculées en θ_0. Le risque statistique est $\frac{1}{2}n^{-1}Trace(I^{-1}J)$ et peut être estimé par $\frac{p}{2n}$ en supposant que le modèle ne soit pas trop "mal spécifié" (voir section 2.4 dans [22]). A noter que dans le cas de modèle bien spécifié $\mathrm{E}_{f^*}\left[\log \frac{f^*(W)}{g^{\theta_0}(W)}\right] = I(g^{\theta_0}, f^*) = 0$ et nous avons $\mathrm{EKL}(g^{\hat{\theta}}; f^*) = \frac{p}{2n} + o(n^{-1})$. Il est aussi possible d'obtenir un développement asymptotique du terme de mauvaise spécification :

$$I(g^{\theta_0}, f^*) = \mathrm{E}_{f^*}\left[\log f^*(W)\right] - n^{-1}\mathrm{E}_{f^*}\left[L^{\hat{\theta}}_{\mathscr{W}_n}\right] + \frac{1}{2}n^{-1}Trace(I^{-1}J) + o(n^{-1}), \tag{2}$$

où $L^{\hat{\theta}}_{\mathscr{W}_n} = \sum_{i=1}^n \log g^{\hat{\theta}}(W_i)$. Malgré cette approximation, il est impossible de pouvoir proposer un estimateur du risque de mauvaise spécification puisque f^* est inconnue. Le premier terme qui est l'opposé de l'entropie de f^* ne peut en général pas être estimé.

Cet exemple montre la difficulté à "estimer" le risque EKL pour un estimateur quelconque $g^{\hat{\theta}}$ (nous mettons estimer entre guillemets car $\mathrm{EKL}(g^{\hat{\theta}}, f^*)$ dépend de n). En combinant l'équation (1) et (2), il vient :

$$\mathrm{EKL}(g^{\hat{\theta}}; f^*) = -H(f^*) + n^{-1}\left[-L^{\hat{\theta}}_{\mathscr{W}_n} + \mathrm{Trace}(I^{-1}J)\right] + o(n^{-1})$$

 ISSN: 2102-6238

où $H(f^*) = -\mathrm{E}_{f^*}[\log f^*(W)]$. Ainsi, le critère d'information d'Akaike défini par $\mathrm{AIC}(g^{\hat{\theta}}) = -2L^{\hat{\theta}}_{\mathscr{W}_n} + 2p$ (obtenu en supposant le modèle bien-spécifié ; ainsi $\mathrm{Trace}(I^{-1}J) = p$) peut être considéré comme un estimateur à une constante près du risque de Kullback-Leibler puisque

$$\mathrm{E}_{f^*}[\mathrm{AIC}] \simeq 2n\left\{\mathrm{EKL}(g^{\hat{\theta}}, f^*) + H(f^*)\right\}.$$

Cependant, un inconvénient du critère d'Akaike est qu'une valeur de AIC ne représente pas un risque de Kullback-Leibler. Les différences de AIC même si elle s'apparentent (à une constante multiplicative près) à des différences de risques de Kullback-Leibler ne sont pas encore interprétables. De plus, la variabilité du critère AIC est souvent ignorée en pratique.

2.3. Différence de risques de Kullback-Leibler

Considérons un autre modèle $(h^{\gamma}) = (h^{\gamma})_{\gamma\in\Gamma}$ et un estimateur $h^{\hat{\gamma}}$, de f^*, défini par une procédure d'estimation basée sur les observations $\mathscr{W}_n$. Le but est d'évaluer le risque associé à chaque estimateur $g^{\hat{\theta}}$ et $h^{\hat{\gamma}}$ afin de choisir l'estimateur ayant le risque le plus faible. Dans cet optique, nous nous intéressons à la différence des risques de Kullback-Leibler :

$$\Delta(g^{\hat{\theta}}, h^{\hat{\gamma}}) = \mathrm{EKL}(g^{\hat{\theta}}, f^*) - \mathrm{EKL}(h^{\hat{\gamma}}, f^*) = \mathrm{E}_{f^*}\left[\log\frac{h^{\hat{\gamma}}(W)}{g^{\hat{\theta}}(W)}\right].$$

Si $\Delta(g^{\hat{\theta}}, h^{\hat{\gamma}}) < 0$, l'estimateur $g^{\hat{\theta}}$ est meilleur que l'estimateur $h^{\hat{\gamma}}$. L'intérêt de considérer une différence de risques réside dans le fait que $\mathrm{E}_{f^*}\left[\log\frac{h^{\hat{\gamma}}(W)}{g^{\hat{\theta}}(W)}\right]$ dépend de f^* seulement par l'espérance alors que le risque associé à chaque estimateur $\mathrm{E}_{f^*}\left[\log\frac{f^*(W)}{g^{\hat{\theta}}(W)}\right]$ dépend doublement de f^* (par $f^*(W)$ et par l'espérance).

2.3.1. Décomposition de la différence de risques

La différence de risques peut être aussi décomposée comme la somme d'une différence de risques de mauvaise-spécification (MSpR : misspecification risk) et d'une différence de risques statistique (StR : Statistical risks) :

$$\Delta(g^{\hat{\theta}}, h^{\hat{\gamma}}) = \Delta_{\mathrm{MSpR}}(g^{\hat{\theta}}, h^{\hat{\gamma}}) + \Delta_{\mathrm{StR}}(g^{\hat{\theta}}, h^{\hat{\gamma}}),$$

où $\Delta_{\mathrm{MSpR}}(g^{\hat{\theta}}, h^{\hat{\gamma}}) = \mathrm{E}_{f^*}\left[\log\frac{h^{\gamma_0}(W)}{g^{\theta_0}(W)}\right]$ et $\Delta_{\mathrm{StR}}(g^{\hat{\theta}}, h^{\hat{\gamma}}) = \mathrm{E}_{f^*}\left[\log\frac{g^{\theta_0}(W)}{g^{\hat{\theta}}(W)}\right] - \mathrm{E}_{f^*}\left[\log\frac{h^{\gamma_0}(W)}{h^{\hat{\gamma}}(W)}\right]$. La différence de risques de mauvaise spécification ne dépend pas de n ni de la procédure d'estimation, contrairement à la différence de risques statistique. Ainsi cette décomposition permet de souligner que nous sommes face à un problème de sélection d'estimateurs et non à un problème de choix de modèle. L'estimateur $g^{\hat{\theta}}$ peut être préféré (au sens du risque de Kullback-Leibler) à $g^{\hat{\gamma}}$ pour une quantité d'information n alors que pour une quantité d'information différente l'estimateur $g^{\hat{\gamma}}$ pourrait être choisi.

 ISSN: 2102-6238

2.3.2. *Interprétation des différences de risques*

Un point important est de pouvoir comparer et d'indiquer si une différence observée en terme de risque de Kullback-Leibler est grande ou petite. Dans [10], il est proposé un guide pour le praticien afin de qualifier la différence obtenue entre deux estimateurs. L'idée était de relier le risque de Kullback-Leibler avec une quantité courante plus interprétable. L'exemple présenté par Commenges et al. [10] concerne l'erreur relative faite par $P_g(A)$ pour évaluer la probabilité d'un événement A basée sur la densité g alors que la vraie densité de probabilité f^* donne $P^*(A)$. L'erreur relative est alors définie par

$$r_e(P_g(A), P_{f^*}(A)) = \frac{P_{f^*}(A) - P_g(A)}{P_{f^*}(A)}.$$

Considérons le cas où l'événement A est l'événement le plus sous-évalué par $P_g(A)$ (au sens ou $\mid P_g(A) - P^*(A) \mid > \mid P_g(A') - P^*(A) \mid, \quad \forall A'$) : il est défini par $A = \{x : g(x) < f^*(x)\}$. Afin d'obtenir une correspondance simple entre le risque de Kullback-Leibler et l'erreur relative nous considérons le cas où $P_{f^*}(A) = 1/2$ et g/f^* est constant sur A et sur A^c. Dans ce cas, il découle l'approximation suivante pour des petites valeurs de $I(g, f^*)$:

$$r_e(P_g(A), P_{f^*}(A)) = \sqrt{1 - e^{-2I(g,f^*)}} \approx \sqrt{2I(g, f^*)}.$$

A des risques de Kullback-Leibler de l'ordre de 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, correspondent des erreurs relatives $r_e(P_g(A), P_{f^*}(A))$ égales à 0.44, 0.14, 0.045 et 0.014. Ces erreurs peuvent être qualifiées de "grandes", "modérées", "petites" et "négligeables". Plusieurs exemples d'interprétation des risques de Kullback-Leibler sont aussi présentés dans [11].

2.3.3. *Extension au modèle de régression*

En régression, l'intérêt se porte généralement sur la distribution conditionnelle de $Y|X = x$ spécifiée, par exemple, par les densités conditionnelles $\left(g^{\theta}_{Y|X}(\cdot|\cdot)\right)_{\theta\in\Theta}$. Vuong [35] propose de définir la divergence de Kullback-Leibler à partir des densités conditionnelles : $KL(g^{\theta}_{Y|X}, f^*_{Y|X}) = \mathrm{E}_{f^*_{Y,X}}\left[\log \frac{f^*_{Y|X}(Y|X)}{g^{\theta}_{Y|X}(Y|X)}\right]$ où le couple (Y,X) est considéré comme une nouvelle observation provenant de la vraie distribution $f^*_{Y,X}$. Cette définition de la divergence de Kullback-Leibler pour des modèles de régression est cohérente avec l'approche proposée dans [10] (appelée "modèle réduit"). Considérons un échantillon de variables i.i.d $(Y_i, X_i), i = 1, \ldots, n$ provenant de $f^*_{Y,X}(y,x) = f^*_{Y|X}(y|x) f^*_X(x)$. Considérons maintenant le modèle $\left(g^{\theta}_{Y,X}(\cdot,\cdot)\right)_{\theta\in\Theta}$ tel que $g^{\theta}_{Y,X}(y,x) = g^{\theta}_{Y|X}(y|x) f^*_X(x)$; le modèle est dit "model réduit" du fait que $f^*_X(\cdot)$ n'est pas modélisée et est supposée connue. La divergence de Kullback-Leibler s'écrit alors :

$$I(g^{\theta}_{Y,X}; f^*_{Y,X}) = \mathrm{E}_{f^*_{Y,X}}\left[\log \frac{f^*_{Y,X}(Y,X)}{g^{\theta}_{Y,X}(Y,X)}\right] = \mathrm{E}_{f^*_{Y,X}}\left[\log \frac{f^*_{Y|X}(Y|X)}{g^{\theta}_{Y|X}(Y|X)}\right].$$

Le terme $f^*_X(\cdot)$ disparaît et nous retrouvons la définition proposée par Vuong. L'approche "modèle réduit" a l'avantage de ne pas nécessiter la définition d'un nouveau critère.

 ISSN: 2102-6238

3. Critères pratiques de sélection : AIC, LCV et EIC

Nous présentons trois critères de sélection permettant d'estimer une différence de risques de Kullback-Leibler : le critère AIC défini dans un contexte paramétrique, la validation croisée appliquée à la vraisemblance (LCV) et un critère bootstrap (EIC) utilisés dans des contextes paramétrique, semi-paramétrique et non-paramétrique. Dans la suite, il sera noté "estimateur" entre guillemets pour indiquer que la quantité d'intérêt dépend de n.

3.1. Différence de AIC

Dans le cas de modèles paramétriques et en considérant des estimateurs par maximum de vraisemblance, Commenges et al. [11] proposent à partir des équations (1) et (2) un "estimateur" de $\Delta(g^{\hat{\theta}}, h^{\hat{\gamma}})$:

$$D(g^{\hat{\theta}}, h^{\hat{\gamma}}) = n^{-1}\left[L^{\hat{\gamma}}_{\mathscr{W}_n} - L^{\hat{\theta}}_{\mathscr{W}_n}\right] + n^{-1}(p-q),$$

où q est la dimension du modèle (h^{γ}). Cet estimateur a un bais en $o(n^{-1})$ et s'apparente à une différence de AIC normalisée. Commenges et al. [11] étudient la variabilité de ce critère en proposant un "intervalle de poursuite" pour $\Delta(g^{\hat{\theta}}, h^{\hat{\gamma}})$. Le nom "'intervalle de poursuite" vient du fait que l'intervalle proposé n'est pas un intervalle de confiance puisque la différence d'intérêt $\Delta(g^{\hat{\theta}}, h^{\hat{\gamma}})$ dépend de n. Par exemple, dans le contexte où $g^{\theta_0} \neq h^{\gamma_0}$, Commenges et al. [11] montrent à partir des résultats de Vuong (1989) que

$$n^{1/2}[D(g^{\hat{\theta}}, h^{\hat{\gamma}}) - \Delta(g^{\hat{\theta}}, h^{\hat{\gamma}})] \xrightarrow[n\to\infty]{L} \mathscr{N}(0, \omega_*^2)$$

où $\omega_*^2 = \text{var}\left[\log \frac{g^{\theta_0}(W)}{h^{\gamma_0}(W)}\right]$. Un estimateur naturel de ω_*^2 est donné par

$$\hat{\omega}_n^2 = n^{-1}\sum_{i=1}^{n}\left[\log\frac{g^{\hat{\theta}}(W_i)}{h^{\hat{\gamma}}(W_i)}\right]^2 - \left[n^{-1}\sum_{i=1}^{n}\log\frac{g^{\hat{\theta}}(W_i)}{h^{\hat{\gamma}}(W_i)}\right]^2.$$

L'intervalle de poursuite est alors défini par $[A_n; B_n]$ où $A_n = D(g^{\hat{\theta}}, h^{\hat{\gamma}}) - z_{1-\alpha/2}n^{-1/2}\hat{\omega}_n$ et $B_n = D(g^{\hat{\theta}}, h^{\hat{\gamma}}) + z_{1-\alpha/2}n^{-1/2}\hat{\omega}_n$ avec $z_{1-\alpha/2}$ le quantile d'ordre $1-\alpha/2$ d'une loi normale centrée réduite. Cet intervalle a la propriété que $P^*[A_n < \Delta(g^{\hat{\theta}}, h^{\hat{\gamma}_n}) < B_n] \longrightarrow 1-\alpha$.

3.2. Différence de LCV

Nous considérons la validation croisée appliquée à la vraisemblance LCV ("Likelihood Cross-Validation") comme un possible "estimateur" (à une constante près : $-H(f)$) de EKL.

$$\text{LCV}(g^{\hat{\theta}}) = -\frac{1}{n}\sum_{i=1}^{n}\log g^{\hat{\theta}_{-i}}(W_i),$$

où $\hat{\theta}_{-i}$ est l'estimateur obtenu sans l'observation i. Ainsi, nous définissons un estimateur de $\Delta(g^{\hat{\theta}}, h^{\hat{\gamma}})$:

$$\text{D}_{\text{LCV}} = \text{LCV}(g^{\hat{\theta}}) - \text{LCV}(h^{\hat{\gamma}}) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{h^{\hat{\gamma}_{-i}}(W_i)}{g^{\hat{\theta}_{-i}}(W_i)}.$$

 ISSN: 2102-6238

A partir des résultats de Liquet et Commenges [24], nous pouvons aisément montrer que dans un contexte paramétrique le biais de D_{LCV} est en $o(n^{-1})$. De plus, pour des estimateurs du maximum de vraisemblance, Stone [33] a montré l'équivalence asymptotique du LCV avec le critère AIC. Enfin, l'avantage de ce critère réside dans le fait qu'il peut être aussi utilisé pour comparer des estimateurs lisses dans des modèles non-paramétriques. Par exemple, ce critère permet de choisir le paramètre de lissage pour des estimateurs par vraisemblance pénalisée ([23], [26]). En revanche, l'inconvénient de ce critère est son temps de calcul puisqu'il nécessite n maximisations de la vraisemblance dans le cas d'estimateurs du maximum de vraisemblance. Il existe toutefois des approximations requérant seulement une maximisation de la vraisemblance (voir [10] et [28]). Dans le cadre d'estimateurs de maximum de vraisemblance pénalisée (voir section 4.2 pour un exemple d'estimateurs de maximum de vraisemblance pénalisé), l'approximisation du LCV s'écrit :

$$LCV(g^{\hat{\theta}}) \approx -n^{-1}[L^{\hat{\theta}}_{\mathscr{W}_n} - \mathrm{Trace}(H^{-1}_{pl_{\mathscr{W}_n}} H_{L_{\mathscr{W}_n}})],$$

où $H_{pl_{\mathscr{W}_n}}$ et $H_{L_{\mathscr{W}_n}}$ sont respectivement la Hessienne de la log-vraisemblance pénalisée et la Hessienne de la log-vraisemblance. Par analogie avec l'expression du critère AIC, le terme $\mathrm{Trace}(H^{-1}_{pl_{\mathscr{W}_n}} H_{L_{\mathscr{W}_n}})$ s'interprète comme le degré de liberté du modèle.

3.3. Différence de EIC

Le critère EIC ("Extented Information Criterion") a tout d'abord été proposé par Konishi et Kitagawa [18] dans un contexte paramétrique et repris ensuite par Liquet et al. [25] dans un contexte semi-paramétrique. L'idée du EIC est de corriger par bootstrap le biais de la log-vraisemblance, considérée comme un estimateur de la partie informative de l'information de Kullback-Leibler :

$$\mathrm{EIC}(g^{\hat{\theta}}) = -\frac{1}{n} L^{\hat{\theta}}_{\mathscr{W}_n} + \hat{b}_{\mathscr{W}_n}$$

où $\hat{b}_{\mathscr{W}_n} = \frac{1}{B}\sum_{j=1}^{B}\left\{\frac{1}{n}\sum_{i=1}^{n}\log g^{\hat{\theta}_j}(W_i^j) - \frac{1}{n}\sum_{i=1}^{n}\log g^{\hat{\theta}_j}(W_i)\right\}$ avec B le nombre de répliques bootstrap de l'échantillon $\mathscr{W}_n$ et $\hat{\theta}_j$ est l'estimateur obtenu à partir du j-ième échantillon bootstrap $\mathscr{W}_n^j = (W_1^j, \ldots, W_n^j)$, ($W_i^j \sim \hat{F}_n$; $\hat{F}_n$ étant la distribution empirique de W). Ainsi, il est possible d'estimer une différence de risques de Kullback-Leibler entre deux estimateurs $g^{\hat{\theta}}$ et $h^{\hat{\gamma}}$ en utilisant le critère EIC :

$$\mathrm{D}_{\mathrm{EIC}} = \mathrm{EIC}(g^{\hat{\theta}}) - \mathrm{EIC}(h^{\hat{\gamma}}).$$

L'intérêt de ce critère par rapport à une différence de LCV est son temps de calcul dans le cas où il n'existe pas de version approchée de la validation croisée appliquée à la vraisemblance et quand $B < n$.

Commentaire : Les performances des critères AIC, LCV et EIC ont été étudiées et comparées à celles d'autres critères en simulation par Liquet et al. [25]. Le critère EIC obtient les meilleurs performances bien que le LCV donne des résultats très proches sauf dans le cas d'échantillon de faible taille.

 ISSN: 2102-6238

4. Extension à des données censurées

Afin de traiter des données plus complexes, nous proposons une formalisation plus générale de la notion de risque de Kullback-Leibler que celle utilisée dans l'introduction. Considérons un espace mesurable $(\Omega, \mathscr{F})$. Étant donné deux probabilités équivalentes P^1 et P^2, et $\mathscr{X}$ une sous-σ-algèbre de $\mathscr{F}$, la perte en utilisant P^2 à la place de P^1 est mesurée par le log du rapport de vraisemblance $L_{\mathscr{X}}^{P^1/P^2} = \log \frac{dP^1}{dP^2}_{|\mathscr{X}}$. En supposant que l'espérance sous P^1 existe, $I(P^2|P^1;\mathscr{X}) = \mathrm{E}_{P^1}[L_{\mathscr{X}}^{P^1/P^2}]$ est le risque de Kullback-Leibler de P^2 relativement à P^1 sur $\mathscr{X}$.

Exemple 2. *Considérons la variable aléatoire X de fonction de densité f_X^1 et f_X^2 sous P^1 et P^2 respectivement et $\mathscr{X}$ la σ-algèbre engendrée par la variable aléatoire X. Dans ce contexte, nous avons $\frac{dP^1}{dP^2}_{|\mathscr{X}} = \frac{f_X^1(X)}{f_X^2(X)}$ et la divergence de la distribution P^2 relative à P^1 s'écrit :*

$$I(P^2|P^1;\mathscr{X}) = \int \log \frac{f_X^1(x)}{f_X^2(x)} f_X^1(x)dx.$$

On retrouve la définition classique (présentée en section 2) de l'information de Kullback-Leibler.

Remarquons que, dans la définition de la divergence de Kullback-Leibler, nous spécifions la σ-algèbre sur laquelle nous évaluons le risque. Le risque peut être évalué sur une σ-algèbre différente, ce qui donnerait évidemment un résultat différent. Ceci est commode en présence de données incomplètes. Les observations sont représentées par la σ-algèbre $\mathscr{O}$. Nous avons $\mathscr{O} = \mathscr{X}$ pour des données complètes et $\mathscr{O} \subset \mathscr{X}$ en présence de données incomplètes (lorsque le mécanisme d'observation est déterministe). Dans ce cas, il est très difficile d'estimer $I(P^2|P^1;\mathscr{X})$ et il devient plus réaliste d'utiliser $I(P^2|P^1;\mathscr{O}) = \mathrm{E}_{P^1}[L_{\mathscr{O}}^{P^1/P^2}]$. Cette flexibilité a permis, en particulier, de développer des critères de choix d'estimateurs pour des données incomplètes [10]. Nous considérons maintenant, le modèle $(P^\theta) = (P^\theta)_{\theta\in\Theta}$ sur $\mathscr{X}$ et la vraie probabilité P^* inconnue. Le risque de Kullback-Leibler de l'estimateur $P^{\hat{\theta}}$ défini à partir d'un échantillon $\bar{\mathscr{O}}_n = \vee_{i=1}^n \mathscr{O}_i$ d'observations $\mathscr{O}_i$ i.i.d est défini par :

$$\mathrm{EKL}(P^{\hat{\theta}};\mathscr{O}_{n+1}) = \mathrm{E}_{P^*}(L_{\mathscr{O}_{n+1}}^{P^*/P^{\hat{\theta}}}).$$

où $\mathscr{O}_{n+1}$ est une nouvelle observation indépendante des $\mathscr{O}_i$ telle que la restriction de P^* à $\mathscr{O}_{n+1}$, $P^*_{\mathscr{O}_{n+1}}$, soit la même que $P^*_{\mathscr{O}_i}$. A partir de cette définition du risque de Kullback-Leibler, il est possible de définir les différentes notions présentées à la section précédente. Les résultats décrits dans la section 2 perdurent dans ce cadre de travail plus général (voir Liquet et Commenges [24] pour plus de détails).
Dans la suite de cette section, nous présentons plusieurs applications dans le cadre de donnée de survie : choix d'estimateurs lisses pour la fonction de risque, choix entre estimateurs issus d'un modèle stratifié et estimateurs issus d'un modèle à risque proportionnel et choix d'estimateurs dans des modèles markovien et non-markovien.

4.1. Choix d'estimateurs lisses pour la fonction de risque

En analyse de survie et dans certaines situations, il est important de pouvoir estimer de façon lisse la fonction de risque $\lambda(\cdot)$. Plusieurs méthodes d'estimations sont alors envisageables, par exemple

 ISSN: 2102-6238

les méthodes à noyaux [29] ou de vraisemblance pénalisée [17]. On note $\widehat{\lambda}_h^{\mathscr{W}}(\cdot)$ un estimateur lisse de $\lambda(\cdot)$ avec h représentant le paramètre de lissage et $\mathscr{W} = (W_1, \ldots, W_n)$ représentant l'échantillon de n variables aléatoires i.i.d. Considérons, par exemple, des données censurées à droite où $W_i = (\tilde{T}_i, \delta_i)$; $\tilde{T}_i = \min(T_i, C_i)$ et $\delta_i = \mathrm{I}_{[T_i \leq C_i]}$. La variable de censure C_i et la variable d'intérêt T_i sont supposées indépendantes et les couples (C_i, T_i) sont identiquement distribués avec comme distribution $F(\cdot)$ pour T_i et $F_C(\cdot)$ pour C_i. On note f et f_C les densités de probabilité, S et S_C les fonctions de survie de T et C respectivement. L'objectif est de choisir l'estimateur $\widehat{\lambda}_h^{\mathscr{W}}(\cdot)$ (noté aussi $P^{\hat{\lambda}_h}$) qui minimise une estimation du risque de Kullback-Leibler :

$$\begin{aligned} \mathrm{EKL}(P^{\hat{\lambda}_h}; \mathscr{O}_{n+1}) &= \mathrm{E}_{P^*}(L_{\mathscr{O}_{n+1}}^{P^*/P^{\hat{\lambda}_h}}) \\ &= \mathrm{E}_{P^*}(L_{\mathscr{O}_{n+1}}^{P^*/P^0}) - \mathrm{E}_{P^*}(L_{\mathscr{O}_{n+1}}^{P^{\hat{\lambda}_h}/P^0}) \end{aligned} \tag{3}$$

où $\mathscr{O}_{n+1}$ est la σ-algèbre générée par une nouvelle observation $W' = (\tilde{T}', \delta')$ de même loi que W_i ($\mathscr{O}_{n+1} = \sigma(\tilde{T}', \delta')$) et P^0 est une mesure de référence quelconque. En prenant la mesure de Lebesgue comme mesure de référence, le deuxième terme de l'équation (3) devient :

$$\begin{aligned} \mathrm{E}_{P^*}(L_{\mathscr{O}_{n+1}}^{P^{\hat{\lambda}_h}/P^0}) &= \mathrm{E}_{P^*}\left\{\log\left[\widehat{f}_h^{\mathscr{W}}(\tilde{T}')\right\}^{\delta'}\{\widehat{S}_h^{\mathscr{W}}(\tilde{T}')\}^{1-\delta'}\{f_C(\tilde{T}')\}^{1-\delta'}\{S_C(\tilde{T}')\}^{\delta'}\right]\right\} \\ &= \mathrm{E}_{P^*}\left\{\log \mathscr{L}_p^{\widehat{\lambda}_h^{\mathscr{W}}}(W')\right\} + \mathrm{E}_{P^*}\{\phi(f_C, S_C, \mathscr{W}')\} \end{aligned} \tag{4}$$

où $\mathscr{L}_p^{\widehat{\lambda}_h^{\mathscr{W}}}(\mathscr{W}') = \{\widehat{f}_h^{\mathscr{W}}(\tilde{T}')\}^{\delta'}\{\widehat{S}_h^{\mathscr{W}}(\tilde{T}')\}^{1-\delta'}$ est la vraisemblance partielle et $\widehat{f}_h^{\mathscr{W}}(\cdot)$ et $\widehat{S}_h^{\mathscr{W}}(\cdot)$ sont les estimateurs de f et S déduit de $\widehat{\lambda}_h^{\mathscr{W}}(\cdot)$. Puisque le second terme de l'équation (4) ne dépend pas de $\widehat{\lambda}_h^{\mathscr{W}}$, chercher l'estimateur qui minimise $\mathrm{EKL}(P^{\hat{\lambda}_h}; \mathscr{O}_{n+1})$ revient à minimiser le terme $-\mathrm{E}_{P^*}\left\{\log \mathscr{L}_p^{\widehat{\lambda}_h^{\mathscr{W}}}(W')\right\}$. Ceci permet de s'affranchir de l'estimation de la distribution de la censure. Il est à noter qu'en travaillant sur des différences de risques de Kullback-Leibler, le terme correspondant à la censure $\phi(f_C, S_C, \mathscr{W}')$ et le terme $\mathrm{E}_{P^*}(L_{\mathscr{O}_{n+1}}^{P^*/P^0})$ disparaissent. Ainsi, il est possible d'estimer, en utilisant par exemple les critères D_{LCV} où D_{EIC}, une différence de risque entre deux estimateurs lisses de la fonction de risque. Dans [23], plusieurs estimateurs du terme $-\mathrm{E}_{P^*}\left\{\log \mathscr{L}_p^{\widehat{\lambda}_h^{\mathscr{W}}}(W')\right\}$ sont proposés pour sélectionner le paramètre de lissage pour les estimateurs par vraisemblance pénalisée et pour des estimateurs à noyaux de la fonction de risque. Finalement, nous pouvons choisir entre ces deux méthodes d'estimation en déterminant la difference de risques entre les deux estimateurs retenus dans les deux approches.

Commentaire : Dans [23], les performances des critères pratiques basées sur le risque de Kullbak-Leibler pour le choix du paramètre de lissage pour les estimateurs à noyaux sont aussi comparées en terme d'erreur quadratique moyenne intégrée (MISE). L'etude de simulation dans [23] montre que la méthode proposée par Ramlau-Hansen [29] (basée sur le MISE) pour choisir le paramètre de lissage obtient de moins bons résultats même en terme de MISE.

4.2. Choix entre un modèle à risques proportionnels et un modèle stratifié

Le risque de Kullback-Leibler peut aussi être utilisé pour choisir entre des modélisations différentes de la fonction de risque. Ce problème se pose dans l'exemple de la modélisation du

 ISSN: 2102-6238

risque de démence cité en introduction. Considérons ici un exemple simple de données censurées à droite (comme défini en 4.1) et l'observation de $\mathbf{X} = (X_1, \ldots, X_n)$ le vecteur d'une variable explicative binaire (codée 0/1). Finalement, nous observons l'échantillon $\mathscr{W} = (W_1, \ldots, W_n)$ avec $W_i = (\tilde{T}_i, \delta_i, X_i)$. Une première modélisation proposée est un modèle à risques proportionnels :

$$\lambda(t|X_i) = \lambda_0(t) \exp \beta X_i \quad i = 1, \ldots, n.$$

Une alternative à cette approche est de considérer un modèle stratifié :

$$\lambda(t|X_i) = \begin{cases} \lambda_0(t) & \text{si } X_i = 0, \\ \lambda_1(t) & \text{si } X_i = 1. \end{cases}$$

Afin de choisir entre ses deux approches, Liquet et al. [26] proposent de juger par le risque de Kullback-Leibler la qualité des estimateurs des deux modèles. Il est à noté que le modèle à risques proportionnels est un sous-modèle du modèle stratifié. Dans le but d'obtenir des estimations lisses de la fonction de risque, les deux modèles sont estimés en utilisant l'approche par vraisemblance pénalisée. Dans le modèle à risques proportionnels, $\widehat{\lambda}_0^h(\cdot)$ et $\widehat{\beta}$ maximisent la vraisemblance pénalisée :

$$p\mathscr{L}_h(\mathscr{W}) \quad = \quad \log \mathscr{L}_p^{\lambda_0,\beta}(\mathscr{W}) - h \int_0^\infty {\lambda_0}^{''2}(u) du$$

où $\log \mathscr{L}_p^{\lambda_0,\beta}(\mathscr{W}) = \sum_{i=1}^n \left[\delta_i \{ \log(\lambda_0(\tilde{T}_i)) + x_i\beta \} - \int_0^{\tilde{T}_i} \lambda_0(u) \exp(x_i\beta) du \right]$ est la log-vraisemblance (conditionnelle aux $X_i, i = 1, \ldots, n$), h représente le paramètre de lissage permettant de contrôler le compromis entre l'adéquation aux données et le lissage de la fonction de risque. La solution de ce problème de maximisation est approchée par une base de splines ([17]).
Dans le modèle stratifié, $\widehat{\lambda}_0^h(\cdot)$ and $\widehat{\lambda}_1^h(\cdot)$ maximisent :

$$p\mathscr{L}_h(\mathscr{W}) \quad = \quad \log \mathscr{L}_p^{\lambda_0,0}(\mathscr{W}^0) + \log \mathscr{L}_p^{\lambda_1,0}(\mathscr{W}^1) - h \int_0^\infty \left\{ {\lambda^0}^{''2}(u) + {\lambda^1}^{''2}(u) \right\} du$$

où $\mathscr{W}^0 = (W_1^0, \ldots, W_{n_0}^0)$ avec $W_i^0 = (\tilde{T}_i, \delta_i, X_i = 0)$ et $\mathscr{W}^1 = (W_1^1, \ldots, W_{n_1}^1)$ avec $W_i^1 = (\tilde{T}_i, \delta_i, X_i = 1)$. Les fonctions de risque $\lambda_0(\cdot)$ et $\lambda_1(\cdot)$ ne sont pas estimées séparément. Ces dernières sont estimées en utilisant le même paramètre de lissage. Ainsi, dans chaque modèle, la méthode du maximum de vraisemblance pénalisée définit une famille d'estimateurs semi-paramétriques indexée par un seul hyper-paramètre h (le paramètre de lissage). Liquet et al. [26] proposent la validation croisée appliquée à la vraisemblance (LCV) ou une version approchée pour sélectionner le paramètre de lissage h dans ces deux familles d'estimateurs. Le critère LCV est ici définie par :

$$\mathrm{LCV}(P^{\hat{\lambda}_h}) = -\frac{1}{n} \sum_{i=1}^n \log \mathscr{L}^{\widehat{\lambda}_h^{\mathscr{W}^{-i}}}(W_i)$$

où $\mathscr{L}^{\widehat{\lambda}_h^{\mathscr{W}^{-i}}}(W_i)$ est la contribution à la vraisemblance de W_i pour l'estimateur défini à partir de l'échantillon $\mathscr{W}^{-i}$ dans lequel l'observation W_i a été enlevée.

Finalement, une différence de LCV permet d'estimer une différence de risques de Kullback-Leibler entre les deux estimateurs retenus et donc de choisir entre ces deux approches.

 ISSN: 2102-6238

Retour sur l'exemple d'application de modélisation du risque de démence : Dans [26], Liquet et al. ont calculé le LCV pour l'estimateur basé sur un modèle stratifié suivant le sexe (LCV=0.4129) et pour un estimateur issu d'un modèle a risques proportionnels (LCV=0.4136). Ce résultat est en faveur d'un estimateur basé sur un modèle stratifié. Nous pouvons alors estimer une différence de risques de Kullback-Leibler par une différence de LCV. Ceci donne une différence de 0.0007 qualifiée de "petite" (d'ordre 10^{-3}) . Par ailleurs, dans [26], plusieurs modélisations tenant compte du niveau d'éducation avait ensuite été effectuées. Le meilleur estimateur obtenu au sens du risque Kullback-Leibler (LCV=0.4071) est issu d'un modèle stratifié sur le sexe et à risque proportionnels sur le niveau d'éducation ; le coefficient de proportionnalité associé au niveau d'étude est le même chez les femmes et les hommes. Ce modèle s'écrit :

$$\lambda(t|S_i,E_i) = \begin{cases} \lambda_0(t)\exp\beta E_i & \text{si } S_i = 0 \text{ (femme)}, \\ \lambda_1(t)\exp\beta E_i & \text{si } S_i = 1 \text{ (homme)}. \end{cases}$$

La différence de risque de Kullback-Leibler de l'estimateur issu de ce modèle par rapport à l'estimateur basé sur le modèle stratifié sur le sexe (sans tenir compte du niveau d'étude) est estimée à 0.4071-0.4136=-0.0065 ; difference de risque qualifiée de modérée (d'ordre 10^{-2}). Il est donc souhaitable de considérer le niveau d'éducation pour estimer le risque de démence. Les sujets n'ayant pas obtenus le certificat d'étude ont un risque plus important de devenir dément. L'estimation par vraisemblance pénalisée de ce dernier modèle fournit une estimation du risque relatif du niveau d'éducation à 1.97 ; l'intervalle de confiance correspondant est [1.63 ; 2.37] (voir [26] pour plus de détail sur cette application).

4.3. Choix d'estimateurs dans des modèles multi-états

Nous présentons, dans cette sous-section, une application du risque de Kullback-Leibler et de son critère estimé le LCV pour le choix d'estimateurs semi-paramétriques dans des modèles multi-états. Nous renvoyons le lecteur au travail de Commenges et al. [10] pour plus de précisions sur ce sujet. Considérons par exemple le modèle "sain-malade-mort" présenté en figure 2. L'état

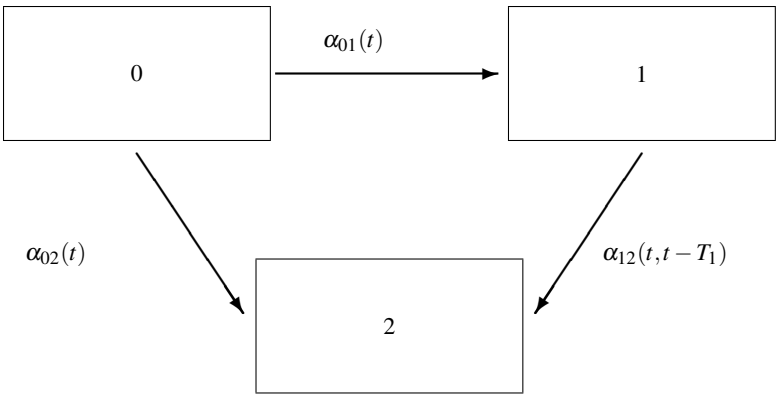

FIGURE 2. *Le modèle 'illness-death"*

 ISSN: 2102-6238

0 représente l'état sain, l'état 1 représente l'état malade et l'état 2 représente l'état décédé. Le modèle multi-états peut être défini par les intensités de transitions $\alpha_{01}(t), \alpha_{02}(t)$ et $\alpha_{12}(t, t-T_1)$ où T_1 représente le temps d'entrée dans l'état 1. Dans ce modèle Illness-death, pour la transition de l'état 1 à l'état 2 on peut hésiter entre plusieurs modèles :

- Un modèle de markov non homogène : $\alpha_{12}(t, t-T_1) = \alpha_{12}(t)$, où $t-T_1$ est le temps passé dans l'état 1.
- Un modèle semi-markov classique : $\alpha_{12}(t, t-T_1) = \alpha_{12}(t-T_1)$ où l'intensité α_{12} dépend du temps passé dans l'état malade.
- Ou d'autres types de modèles semi-markoviens, en supposant par exemple un risque de décès plus élevé pour les sujets de l'état 1 par rapport au sujet de l'état 0 et qui dépend du temps passé dans l'état 1 : $\alpha_{12}(t, t-T_1) = \alpha_{02}(t) + \alpha_{12}(t-T_1)$.

Il est aussi envisageable, pour chacun de ces modèles, d'incorporer l'information de variables explicatives $Z_i(t)$ de l'individu i de façon additive (dans l'idée du modèle additif de Aalen [1]) ou de façon multiplicative (dans l'idée du modèle à risques proportionnels). Par exemple, une structure multiplicative pour les variables explicatives est :

$$\alpha^i_{01}(t) = \alpha^0_{01}(t)\exp(\beta_{01}Z_i(t)),$$

$$\alpha^i_{02}(t) = \alpha^0_{02}(t)\exp(\beta_{02}Z_i(t)),$$

$$\alpha^i_{12}(t, t-T_1) = \alpha^0_{12}(t, t-T_1)\exp(\beta_{12}Z_i(t)),$$

où α_{01}, α_{02} et α_{12} sont les fonctions de risque de base. Des estimateurs semi-paramétriques obtenus par la méthode de la vraisemblance pénalisée peuvent être définis pour chacun de ces modèles (définis par leurs hypothèses markoviennes ou non-markoviennes et leurs structures additives ou multiplicatives). L'objectif est alors de choisir le meilleur estimateur au sens du EKL pour représenter au mieux les données observées. Enfin, il est aussi possible en travaillant sur les différences de risques de "quantifier" une mesure de proximité entre les estimateurs. Dans [10], les modèles Multi-états considérés sont présentés par des processus de comptage multivariés ([9]). Un des avantages de cette représentation des modèles multi-états est la possibilité d'utiliser la formule générale du rapport de vraisemblance de Jacod ([16]) qui intervient dans la définition du EKL et par suite dans la definition du LCV. Il est proposé dans [10] une version approchée du LCV pour choisir entre les différents estimateurs semi-paramétriques.

5. Choix d'estimateurs basés sur des observations différentes

Il est assez courant en épidémiologie que nous souhaitions évaluer la qualité des estimateurs sur un ensemble particulier d'informations, tandis que les estimateurs sont définis sur un ensemble plus large d'informations. Dans ce contexte, Liquet et Commenges [24] proposent de définir un critère, basé sur le risque de Kullback-Leibler, permettant de comparer des estimateurs définis à partir d'observations différentes. Les estimateurs considérés sont alors évalués sur leur information commune. Dans cette section, une synthèse des principaux résultats obtenus dans [24] est présentée.

 ISSN: 2102-6238

5.1. *Exemples d'application*

5.1.1. *Modèle linéaire versus modèle binaire*

Le premier exemple concerne la prédiction d'un événement du type $\{Y > l\}$ à partir de plusieurs variables explicatives. La variable Y peut représenter, par exemple, le taux de cholestérol et l'événement $\{Y > l = 220\ mg/dl\}$ indiquer une hypercholestérolémie. L'objectif est de prédire une hyperchoslestérolémie sachant un certain nombre de facteurs de risque. Une première approche est de modéliser l'hypercholestérolémie en utilisant un modèle de régression binaire sur la variable $Z = 1_{\{Y>l\}}$. Des modèles de type "probit" ou "logit" sont alors envisageables. Ces modèles sont définis par :

$$P(Z = 1|X) = F(X^T\beta),$$

où β est un vecteur de paramètre associé au vecteur de variables explicatives X et F est une fonction de répartition connue. A partir de l'estimation de ce modèle, il est possible d'estimer pour un individu, la probabilité d'être en hypercholestérolémie connaissant ses variables explicatives. Une alternative est de modéliser la variable continue Y par un modèle de régression linéaire :

$$Y = X^T\beta + \sigma\varepsilon,$$

où l'erreur ε est supposé centrée, de variance unité et de distribution connue F_ε. A partir de l'estimation de ce modèle, la prédiction de l'événement d'intérêt est définie par $F_\varepsilon(\frac{X^T\widehat{\beta}-l}{\widehat{\sigma}})$ (en supposant que la densité de ε est symétrique). Cette dernière approche utilise plus d'information mais peut nécessiter plus d'hypothèses et plus de paramètres. Le problème est alors de comparer ces deux approches. Pour cela, nous avons besoin d'évaluer la capacité de prédiction du modèle de régression linéaire sur une information plus petite que celle qui est utilisée pour estimer le modèle. Nous renvoyons le lecteur à Liquet et Commenges [24] pour plus de détails.

5.1.2. *Modèle de survie versus modèle Multi-états*

Dans ce deuxième exemple, nous considérons un modèle de survie et nous disposons d'information supplémentaire sur un autre événement qui peut modifier le risque de l'événement d'intérêt. Le modèle multi-états "Illness-death" permet de modéliser les deux événements et permet d'estimer de façon précise la survenue de l'événement d'intérêt. Il est proposé dans [24] de comparer ces deux approches à partir de l'information commune utilisée, c'est à dire la survie de l'événement d'intérêt.

5.2. *Extension du risque de Kullback-Leibler*

Dans le cas classique présenté en section 3, la différence de risques est évaluée sur $\mathscr{O}_{n+1}$, une nouvelle observation indépendante des observations $\mathscr{O}_i$ sur lesquelles les estimateurs en comparaison sont définis. Dans les deux exemples exposés en section 5.1., les estimateurs ne sont pas définis sur les mêmes observations. Dans le premier exemple, le modèle de régression linéaire, noté (P^θ), est estimé à partir des observations $\mathscr{O}_i = \sigma(Y_i, X_i)$ où X_i est un vecteur de variables explicatives de l'individu i. En revanche, le modèle binaire, noté (P^γ), est estimé à partir des observations

 ISSN: 2102-6238

$\mathscr{O}_i^{'} = \sigma(Z_i, X_i)$ où $Z_i = 1_{\{Y_i > l\}}$. Nous avons donc que $\mathscr{O}_i^{'} \subset \mathscr{O}_i$. Ainsi, nous proposons dans [24] une extension du critère classique en évaluant le risque sur une sous σ-algèbre $\mathscr{O}_{n+1}^{'}$ de $\mathscr{O}_{n+1}$ ($\mathscr{O}_{n+1}^{'} \subset \mathscr{O}_{n+1}$). Le risque de $P^{\hat{\theta}}$ est évalué sur $\mathscr{O}_{n+1}^{'}$ par :

$$\begin{aligned} \mathrm{EKL}(P^{\hat{\theta}}; \mathscr{O}'_{n+1}) &= \mathrm{E}_{P^*}(L^{P^*/P^{\hat{\theta}}}_{\mathscr{O}'_{n+1}}) \\ &= \mathrm{E}_{P^*}(L^{P^*/P^0}_{\mathscr{O}'_{n+1}}) - \mathrm{E}_{P^*}(L^{P^{\hat{\theta}}/P^0}_{\mathscr{O}'_{n+1}}), \end{aligned}$$

où P^0 est une mesure de référence quelconque.

Remarque 1. *Dans le premier exemple, en prenant la mesure de Lebesgue comme mesure de référence, on obtient simplement que*

$$L^{P^{\hat{\theta}}/P^0}_{\mathscr{O}'_{n+1}} = Z_{n+1} \log \hat{p}_{n+1} + (1 - Z_{n+1}) \log(1 - \hat{p}_{n+1}),$$

où $\mathscr{O}_{n+1}^{'} = \sigma(Z_{n+1}, X_{n+1})$ *et* $\hat{p}_{n+1} = P^{\hat{\theta}}(Z_{n+1} = 1 | X_{n+1})$ *est déduit de l'estimateur du modèle de régression linéaire* (P^{θ}).

La différence de risques entre les deux estimateurs $P^{\hat{\theta}}$ et $P^{\hat{\gamma}}$ est alors définie par :

$$\Delta(P^{\hat{\theta}}, P^{\hat{\gamma}}; \mathscr{O}_{n+1}^{'}) = \mathrm{EKL}(P^{\hat{\theta}}; \mathscr{O}'_{n+1}) - \mathrm{EKL}(P^{\hat{\gamma}}; \mathscr{O}'_{n+1}) = \mathrm{E}_{P^*}(L^{P^{\hat{\gamma}}/P^{\hat{\theta}}}_{\mathscr{O}'_{n+1}}).$$

5.3. Critère de sélection pratique

Trois critères pratiques, $\mathrm{D}_{\mathrm{RStR}}$, $\mathrm{D}_{\mathrm{RLCV}}$ et $\mathrm{D}_{\mathrm{RAIC}}$ ont été développés pour estimer une différence de risques de Kullback-Leibler.

5.3.1. Modèles bien spécifiés

Dans le cadre de modèles bien spécifiés, la différence de risques de mauvaise-spécification devient nulle et la différence de risques de Kullback-Leibler est égale à la différence de risques statistique. Dans ce cadre et pour des estimateurs du maximum de vraisemblance, un estimateur de $\Delta(P^{\hat{\theta}}, P^{\hat{\gamma}}; \mathscr{O}_{n+1}^{'})$ est défini par :

$$\mathrm{D}_{\mathrm{RStR}} = (2n)^{-1} \left[\mathrm{Trace} \left\{ \widehat{\mathbb{I}}^{\mathscr{O}'_{n+1}}_{P_{\theta_0}} (\widehat{\mathbb{I}}^{\mathscr{O}_{n+1}}_{P_{\theta_0}})^{-1} \right\} - p \right],$$

où $\widehat{\mathbb{I}}^{\mathscr{O}'_{n+1}}_{P_{\theta_0}} = -\frac{1}{n} \frac{\partial^2 L^{P^{\theta}/P^*}_{\bar{\mathscr{O}}'_n}}{\partial \theta^2}|_{\hat{\theta}_n}$, $\widehat{\mathbb{I}}^{\mathscr{O}_{n+1}}_{P_{\theta_0}} = -\frac{1}{n} \frac{\partial^2 L^{P^{\theta}/P^*}_{\bar{\mathscr{O}}_n}}{\partial \theta^2}|_{\hat{\theta}_n}$ avec $\bar{\mathscr{O}}'_n = \vee_{1 \le i \le n} \mathscr{O}_i^{'}$ et p est le nombre de paramètre du modèle (P^{γ}) défini sur la petite σ-algèbre. Ici RStR signifie "restricted statistical risk". Il est montré dans [24] que cet estimateur a un biais en $o(n^{-1})$.

 ISSN: 2102-6238

5.3.2. *Modèles mal spécifiés*

Dans le cadre de modèles mal spécifiés et pour des estimateurs du maximum de vraisemblance, le $\mathrm{D}_{\mathrm{RAIC}}$ (pour "restricted" AIC) défini par :

$$\mathrm{D}_{\mathrm{RAIC}} = n^{-1}\left[L_{\bar{\mathscr{O}}'_n}^{P^{\hat{\gamma}}/P^{\hat{\theta}}} + \mathrm{Trace}\left\{\widehat{\mathbb{I}}_{P_{\theta_0}}^{\mathscr{O}'_{n+1}}\left(\widehat{\mathbb{I}}_{P_{\theta_0}}^{\mathscr{O}_{n+1}}\right)^{-1}\right\} - p\right],$$

est un estimateur de $\Delta(P^{\hat{\theta}}, P^{\hat{\gamma}}; \mathscr{O}'_{n+1})$.

Un second estimateur de $\Delta(P^{\hat{\theta}}, P^{\hat{\gamma}}; \mathscr{O}'_{n+1})$ est défini par une différence de LCV modifié :

$$\mathrm{D}_{\mathrm{RLCV}} = \mathrm{LCV}(P^{\hat{\theta}_n}; \bar{\mathscr{O}}'_n) - \mathrm{LCV}(P^{\hat{\gamma}_n}; \bar{\mathscr{O}}'_n) = \frac{1}{n}\sum_{i=1}^{n} L_{\mathscr{O}'_i}^{P^{\hat{\gamma}_{-i}}/P^{\hat{\theta}_{-i}}},$$

où $\mathrm{LCV}(P^{\hat{\theta}_n}; \bar{\mathscr{O}}'_n) = -\frac{1}{n}\sum_{i=1}^{n} L_{\mathscr{O}'_i}^{P^{\hat{\theta}_{-i}}/P^0}$, avec $\hat{\theta}_{-i} = \hat{\theta}(\bar{\mathscr{O}}_{n|i})$ l'estimateur défini à partir de $\bar{\mathscr{O}}_{n|i} = \vee_{j\neq i}\mathscr{O}_j$, $\hat{\gamma}_{-i} = \hat{\gamma}(\bar{\mathscr{O}}'_{n|i})$ l'estimateur défini à partir de $\bar{\mathscr{O}}'_{n|i} = \vee_{j\neq i}\mathscr{O}'_j$ et P^0 une probabilité de référence. Ici RLCV signifie "restricted" LCV. Dans un contexte paramétrique cet estimateur conduit a un biais en $o(n^{-1})$ pour la différence de risque de Kullback-Leibler. De plus, ce dernier critère peut être utilisé pour comparer des estimateurs semi-paramétriques lisses.

6. Conclusion

Plusieurs critères pratiques de sélection vus sous l'angle commun du risque de Kullback-Leiber ont été présentés dans un cadre standard. Il a été souligné l'intérêt d'estimer des différences de risques de Kullback-Leibler. Une interprétation de ces différences de risques a notamment été exposée. Cependant, un travail supplémentaire est nécessaire pour approfondir l'intuition sur l'ordre de grandeur d'une différence de risque de Kullback-Leibler. Dans des cadres non standards, deux extensions du risque de Kullback-Leibler ont été exposées. La première extension concerne le choix d'estimateurs en présence de données censurées. La seconde concerne la situation où l'un des estimateurs en comparaison est défini sur une quantité d'information plus importante que celle qui est utilisée pour évaluer le risque. Dans le même ordre d'idée, une perspective de travail est de s'intéresser au choix d'estimateurs dans des études pronostiques où des modèles plus ou moins complexes, comme des modèles conjoints, peuvent être définis pour prédire l'occurrence d'un événement.

Références

[1] O. O. Aalen : Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6:701–726, 1978.

[2] H. Akaike : Information theory and an extension of the maximum likelihood principle. *In* B.N. Petrov et F. Csaki, éditeurs : *Second International Symposium on Information Theory*, pages 267–281, Budapset, 1973. Akademiai kiado.

[3] A.R. Barron, L. Birgé et P. Massart : Risk bounds for model selection via penalization. probab. *Probability Theory and related Fields*, 113(3):301–415, 1999.

 ISSN: 2102-6238

[4] L. BIRGÉ et P. MASSART : Minimal penalties for gaussian model selection. *Probability Theory and related Fields*, 138(1-2):33–73, 2007.

[5] H. BOZDOGAN : Akaike's information criterion and recent developments in information complexity. *J. Math. Psych.*, 44:62–91, 2000.

[6] E. BRUNEL et Fabienne. COMTE : Adaptive estimation of hazard rate with censored data. *Communications in Statistics, Theory and methods*, 37(8):1284–1305, 2008.

[7] K. P. BURNHAM et D. R. ANDERSON : *Model selection and multimodel inference : a practical information-theoretic approach*. 2nd Edition. Springer-Verlag, New York, 2002.

[8] G. CLAESKENS et N. L. HJORT : *Model selection and model averaging*. Cambridge University Press, 2008.

[9] D. COMMENGES et A. GEGOUT-PETIT : Likelihood inference for incompletely observed stochastic processes : ignorability conditions. *Scandinavian Journal of Statistics*, 34(2):432–450, 2007.

[10] D. COMMENGES, P. JOLY, A. GEGOUT-PETIT et B. LIQUET : Choice between semi-parametric estimators of markov and non-markov multi-state models from generally coarsened observations. *Scandinavian Journal of Statistics*, 34:33–52, 2007.

[11] D. COMMENGES, A. SAYYAREH, L. LETENNEUR, J. GUEDJ et A. BAR-HEN : Estimating a difference of Kullback-Leibler risks using a normalized difference of AIC. *Annals of Applied Statistics*, 2(3):1123–1142, 2008.

[12] P. CRAVEN et G. WAHBA : Smoothing noisy data with spline functions : estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerical Mathematics*, 31:377–403, 1979.

[13] J. DELEEUW : *Breakthroughs in statistics*, volume 1, chapitre Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle, pages 599–609. Springer-Verlag, London, 1992. Kotz, S. and Johnson, N. L.

[14] A.L. GIBBS et F.E. SU : On choosing and bounding probability metrics. *International statistical review*, 70:419–435, 2002.

[15] T. J. HASTIE et R. J. TIBSHIRANI : *Generalized Additive Models*. Chapman and Hall, London, 1990.

[16] J. JACOD : Multivariate point processes : predictable projection, Radon-Nikodym derivatives, representation of martingales. *Wahrsch. verw Geb*, 31:235–253, 1975.

[17] P. JOLY, D. COMMENGES et L. LETENNEUR : A penalized likelihood approach for arbitrarily censored and truncated data : application to age-specific incidence of dementia. *Biometrics*, 54:185–194, 1998.

[18] S. KONISHI et G. KITAGAWA : Generalised information criteria in model selection. *Biometrika*, 83:875–890, 1996.

[19] S. KONISHI et G. KITAGAWA : *Information Criteria and Statistical Modeling*. New-York : Springer Series in Statistics, 2008.

[20] S. KULLBACK et R. A. LEIBLER : On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[21] L. LETENNEUR, D. COMMENGES, J. DARTIGUES et P. BARBERGER-GATEAU : Incidence of dementia and alzheimer's disease in elderly community residents of south-western france. *International Journal of Epidemiology*, 23:1256–1261, 1994.

[22] H. LINHART et W. ZUCCHINI : *Model selection*. New-York : Wiley, 1986.

[23] B. LIQUET et D. COMMENGES : Estimating the expectation of the log-likelihood with censored data for estimator selection. *Lifetime Data Analysis*, 10:351–367, 2004.

[24] B. LIQUET et D. COMMENGES : Choice of estimators based on different observations : modified AIC and LCV criteria. *Scandinavian Journal of Statistics*, In press, 2010.

[25] B. LIQUET, C. SAKAROVITCH et D. COMMENGES : Bootstrap choice of estimators in non-parametric families : an extension of EIC. *Biometrics*, 59:172–178, 2003.

[26] B. LIQUET, J. SARACCO et D. COMMENGES : Selection between proportional and stratified hazards models based on expected log-likelihood. *Computational Statistics*, 22:619–634, 2007.

[27] P. MASSART : Sélection de modèles : de la théorie à la pratique. *Journal de la SFDS*, 4:5–28, 2008.

[28] F. O'SULLIVAN : Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific Computing*, 9:363–379, 1988.

[29] H. RAMLAU-HANSEN : Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics*, 11:453–466, 1983.

 ISSN: 2102-6238

[30] G. SCHWARZ : Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[31] R. SHIBATA : Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica*, 7:375–394, 1997.

[32] B.W. SILVERMAN : *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.

[33] M. STONE : An asymptotic equivalence of choice of model by cross-validation and akaike's criterion. *Journal of the Royal Statistical Society B*, 39:44–47, 1974.

[34] A. TSYBAKOV : Introduction à l'estimation non-paramétrique. *In Mathématiques and Applications*, volume 41. Springer-Verlag, Berlin, 2004.

[35] Q.H. VUONG : Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, 57:307–333, 1989.

 ISSN: 2102-6238