Sélection de modèles semi-paramétriques

Benoit Liquet

UNIVERSITE VICTOR SEGALEN DE BORDEAUX II

Année 2002 Thèse n° 958

THESE

pour le

DOCTORAT DE L'UNIVERSITE DE BORDEAUX 2

Mention : Sciences Biologiques et Médicales

Option : Epidémiologie et Intervention en Santé Publique

présentée et soutenue publiquement

le 11 décembre 2002

par **Benoit Liquet**

Né le 7 avril 1975 à Saint Quentin

Sélection de modèles semi-paramétriques

Membres du Jury

Monsieur le Professeur	Avner BAR-HEN	Rapporteur
Monsieur le Professeur	Pascal MASSART	Rapporteur
Monsieur le Professeur	Lucien BIRGE	Examinateur
Monsieur le Professeur	Guy THOMAS	Président
Monsieur le Docteur	Daniel COMMENGES	Directeur de thèse

Remerciements

Je tiens tout d'abord à remercier le Professeur Salamon de m'avoir accueilli dans son laboratoire et de m'avoir permis de mener mon travail de recherche dans d'excellentes conditions.

Je remercie très chaleureusement Pascal Massart pour son accueil dans son laboratoire qui m'a permis de m'ouvrir aux méthodes développées par son équipe. Vous me faites un grand honneur en acceptant d'être le rapporteur de cette thèse et je tiens à vous exprimer toute ma reconnaissance.

Je remercie très sincèrement Avner Bar-Hen pour avoir accepté d'être rapporteur de cette thèse. Je lui suis reconnaissant du temps qu'il a consacré à l'examen de ce travail et je le remercie pour ses remarques et conseils. J'apprécie réellement sa présence dans le jury.

J'adresse mes sincères remerciements à Guy Thomas d'avoir accepté de présider le jury de cette thèse. Veuillez trouver ici l'expression de ma reconnaissance et de ma considération respectueuse.

Merci également à Lucien Birgé qui a accepté de faire partie du jury. C'est un honneur et un plaisir de vous voir participer au jury de soutenance.

Je souhaite exprimer ma profonde gratitude à Daniel Commenges qui a accepté de m'encadrer pendant cette thèse. Sa disponibilité, ses conseils et nos longs entretiens m'ont été d'un grand secours dans l'apprentissage du métier de chercheur. Merci encore.

Mes remerciements les plus chaleureux vont aussi à :

Chantal pour la confiance que vous m'avez accordée en me présentant à Daniel Commenges. J'ai tout à fait conscience que c'est grâce à vos précieux conseils que j'ai pu trouver la motivation pour poursuivre mes études dans cette voie.

Mikhail Nikouline ainsi qu'à toute son équipe pour leur accueil durant mon DEA et mon début de thèse.

Luc, Renaud, Seb et Sophie pour leur relecture attentive et constructive. Un grand merci pour tout le temps que vous m'avez consacré et pour tous vos précieux conseils.

Tous les membres de l'Unité INSERM 330 et en particulier l'équipe Biostat, pour leur sympathie et leur aide, notamment Alioum, Alphonse, Anne, Charlotte, Hélène, Julien, Marie-Noëlle, Marthe-Aline, Reza, Rodolphe, Valérie, Virginie. Je décerne une mention particulière à Luc (encore lui !!) et Pierre pour le temps et l'aide efficace qu'ils m'ont accordés.

Tous les joyeux lurons du labo : Anne, Caro, Cécile, Dorothée, Eric, François, Karine, Laurent, Patrick, Majid, Nico, Raphaelle.

Patou, Marco, Lou, Boby, Midrouill, Sutanto, Didou, Vincent, Willy, Clara, Céline, Sophie, Stef, Virginie, mes amis, mes confidents, ceux qui me font rire et qui ont su me soutenir dans tous les moments. Cette thèse est enfin l'occasion de vous exprimer mon profond attachement.

Maman, Papa, Alain, Martine, Célia, Marina, Olivia, et toute ma famille qui ont contribué à leur manière à la réalisation de ce travail ...

Résumé

Cette thèse développe des méthodes de sélection de modèles pour des applications en Biostatistique et plus particulièrement dans le domaine médical.

Dans la première partie, nous proposons une méthode et un programme de correction du niveau de signification d'un test lorsque plusieurs codages d'une variable explicative sont essayés. Ce travail est réalisé dans le cadre d'une régression logistique et appliqué à des données sur la relation entre cholestérol et démence.

La deuxième partie de la thèse est consacrée au développement d'un critère d'information général permettant de sélectionner un estimateur parmi une famille d'estimateurs semi-paramétriques. Le critère que nous proposons est basé sur l'estimation par bootstrap de l'information de Kullback-Leibler. Nous appliquons ensuite ce critère à la modélisation de l'effet de l'amiante sur le risque de mésothéliome et nous comparons cette approche à la méthode de sélection de Birgé-Massart.

Enfin, la troisième partie présente un critère de sélection en présence des données incomplètes. Le critère proposé est une extension du critère developpé dans la deuxième partie. Ce critère, construit sur l'espérance de la log-vraisemblance observée, permet en particulier de sélectionner le paramètre de lissage dans l'estimation lisse de la fonction de risque et de choisir entre des modèles stratifiés et des modèles à risques proportionnels. Nous avons notamment appliqué cette méthode à la modélisation de l'effet du sexe et du niveau d'éducation sur le risque de démence.

Mots-clefs : bootstrap, épidémiologie, information de Kullback-Leibler, lissage, modèles de survie, multiplicité, p-value, régression logistique, semi-paramétrique, test du score, validation croisée, sélection de modèles.

Abstract

This thesis develops model selection method for Biostatistical applications, essentially in medicine. The study is structured into three parts.

In the first part, we propose a method and a program to determine a significance level for a series of codings of an explanatory variable in logistic regression. This method is illustrated using the data of a study of the relation between cholesterol and dementia.

The second part of the thesis presents a general criterion for choosing an estimator in a family of semi-parametric estimators. It is based on a bootstrap estimator of the Kullback-Leibler information. This criterion is applied for modeling the effect of the asbestos on the risk of mesotheliome and we compare it to Birge-Massart selection method.

The last part provides a criterion for choosing an estimator in a family of semi-parametric estimators from incomplete data. It is an extension of the criterion presented in the second part. This criterion is based on the expected observed log-likelihood. It is used in particular to select the smoothing parameter for the hazard function and to choose between stratified and unstratified proportional hazards models. An example is given for modeling the effect of sex and educational level on the risk of developing dementia.

Key-words : bootstrap, cross-validation, epidemiology, Kullback-Leibler information, logistic regression, model selection, multiplicity, p-value, score tests, semi-parametric, smoothing, survival model.

Table des matières

Introduction générale

Devant la multitude des modèles existants, qui n'a jamais été confronté à la difficulté de sélectionner le modèle le plus adapté à ses données ? Une des difficultés intervenant dans la sélection de modèles de régression est celle du choix des variables explicatives. La sélection de modèle peut être envisagée sous différentes formes. Un premier problème concerne la sélection d'un sous ensemble de variables parmi p variables. On peut aussi voir la sélection de modèle comme la meilleure représentation possible de l'effet d'une variable explicative sur une variable dépendante. Enfin la sélection de modèles intervient dans le choix du paramètre de lissage dans les approches non-paramétriques.

Dans des modèles de régression où la variable dépendante peut être expliquée par p variables, de nombreux modèles paramétriques (2^p choix de variables) sont alors disponibles. Choisir le modèle approprié parmi toutes les possibilités est crucial. Dans les modèles de régression linéaire, une des solutions a été d'utiliser le critère C_p proposé par Mallows [27]. Par ailleurs de nombreux auteurs se sont également penchés sur ce problème de sélection pour des modèles de régression plus généraux [6]. Toujours dans un cadre paramétrique, on peut noter le critère d'Akaike [1] (le AIC) basé sur l'espérance de l'information de Kullback-Leibler [26]. Bien que ce critère soit largement utilisé, il présente certaines déficiences [5]. Le AIC peut être en effet biaisé quand il est utilisé pour des problèmes de régressions paramétriques dans des petits échantillons. Le AIC_c proposé par Hurvitch et Tsai [20] cherche à corriger ce biais et fournit des meilleurs choix de modèles pour des échantillons de faible taille. Une approche par bootstrap peut aussi être utilisée pour estimer ce biais. Le critère défini par Ishiguro, Sakamoto et Kitagawa [21] (le EIC : extended information criterion), est construit de cette façon. Shibata [36] a

prouvé une équivalence asymptotique entre ce critère et le AIC. Le EIC a été utilisé dans le problème de sélection de variable et semble être meilleur que les autres critères [30]. Schwartz [33] propose un critère de sélection consistant (le BIC : bayesien information criterion), permettant de sélectionner le vrai modèle, supposé de dimension finie, avec une probabilité approchant 1 dans des grands échantillons. Le fondement du BIC est bayesien et ne constitue donc pas un estimateur de l'information de Kullback-Leibler. Draper [9] et Potscher [31] fournissent une revue récente des méthodes bayesiennes appliquées dans ce domaine.

La validation croisée est considérée comme une méthode de base de la sélection de modèles ([40], [39], [11]). Le principe est de partitionner l'échantillon en deux sous-échantillons. Le premier est utilisé pour estimer le modèle et le second pour le valider (généralement le second contient seulement une observation). Des critères tel que l'erreur de prédiction sont généralement utilisés en répétant ce processus plusieurs fois. Des critères ont été développés sur ce principe et sont très utilisés ([8], [4], [35], [42] , [18]). Cependant ces méthodes requièrent un temps de calcul important, surtout avec des échantillons de grande taille.

Une approche différente pour résoudre le problème de sélection de modèles est d'utiliser les tests d'hypothèses. Les tests séquentiels sont souvent employés ; les procédures pas à pas descendante et pas à pas ascendante sont très populaires. Ils permettent à chaque pas d'accepter ou d'exclure une variable. Ces procédures sont basées sur un choix subjectif du niveau de signification α (généralement $\alpha = 0.05$ ou 0.01) ; cependant Rawling [32] recommande 0.15 dans le cadre de procédure pas à pas appliquée à des modèles de régression. Le problème majeur de ces procédures est la multitude de tests effectués [41] qui ne sont généralement pas indépendants. De plus, les tests entre des modèles qui ne sont pas emboités sont problématiques. Finalement, plusieurs auteurs ([2], [34]) considèrent que les tests d'hypothèses ne sont pas adaptés à la sélection de modèles et souhaiteraient que l'attrait pour ces procédures n'augmentent pas dans les années à venir.

Dans des problèmes non-paramétriques comme l'estimation de l'effet d'une variable de regression [17], l'estimation de la densité [38] ou l'estimation de fonction de risque [29], [22],

le nombre de paramètres est considéré comme infini. Dans le but d'obtenir des estimations lisses, les méthodes d'estimations non pramétriques comme l'approche par vraisemblance pénalisée [37] ou la méthode de lissage à noyau [16] dépendent alors d'un paramètre de lissage. La sélection de modèles (ou plutôt d'estimateurs ou de procédures d'estimations) se résume alors au choix d'une valeur pour le paramètre de lissage. Les critères paramétriques comme le AIC et le AIC_c peuvent être adaptés pour déterminer le paramètre de lissage dans des problèmes simples de régression linéaire [19]. Par ailleurs, la validation croisée basée sur une erreur quadratique (noté CV) est généralement utilisée pour résoudre ce choix [13]. Graven et Wahba [8] proposent le GCV (generalized cross-validation), une version approchée du CV. Cependant plusieurs auteurs soulèvent quelques limites de ces critères ([14], [15]). Green et Silverman [13] construisent alors un critère de validation croisée basé sur la vraisemblance. Ce critère, le LCV (likelihood cross validation), coûteux en temps de calcul, est souvent remplacé par des versions simplifiées et approchées [29].

Finalement, l'épidémiologiste est non seulement confronté à la difficulté de sélectionner un modèle approprié à ses données mais aussi à l'embarras de choisir entre les différents critères de sélection. Les méthodes de choix (tests d'hypothèses, critères bayesien, validation croisée) sont construits avec des esprits différents. Elles se limitent souvent à des questions précises et sont utilisées dans un contexte particulier.

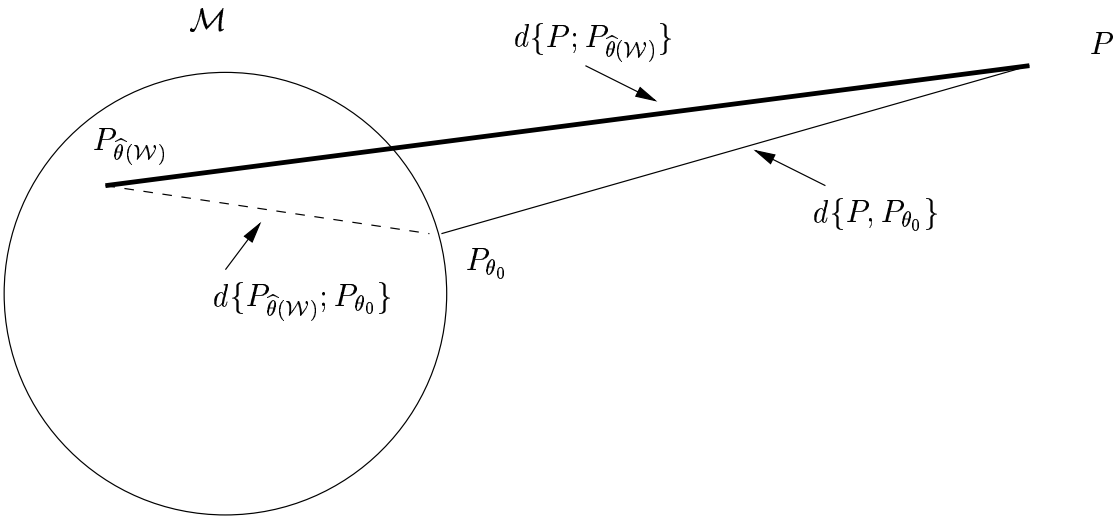
Nous proposons dans cette thèse d'apporter une solution à un problème lié aux procédures des tests d'hypothèses. Dans la recherche du meilleur codage d'une variable explicative, afin de représenter son effet sur la variable dépendante, une série de tests est généralement effectuée. Malheureusement, les praticiens omettent de corriger la P_{value} associée au test le plus significatif. Nous résoudrons ce problème dans le chapitre 1 de la thèse. Ce premier travail peut être assimilé à un problème de sélection de modèles. En effet à un codage correspond un modèle. Le processus de sélection du meilleur codage revient à une procédure de sélection de modèles.

Le chapitre 1 a été l'inspirateur de l'objet principal de la thèse à savoir la sélection de

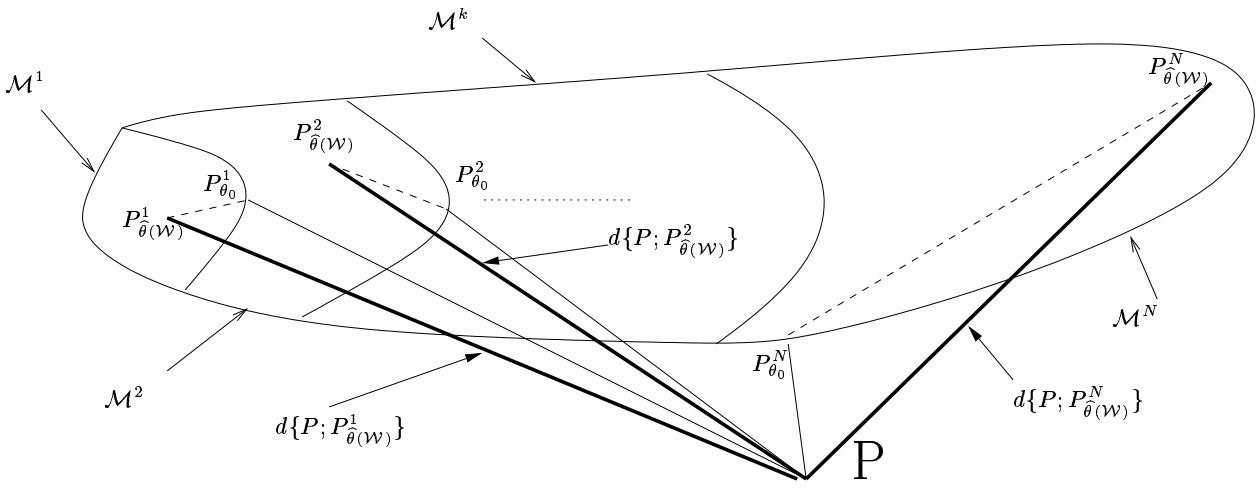
modèles par un critère d'information. Nous proposons dans les autres chapitres d'unifier les différentes approches par un critère d'information répondant au principe de la sélection de modèles à la fois dans un cadre paramétrique et non-paramétrique. Cet unique critère d'information répond aux exigences du principe même de la sélection de modèles que nous présentons.

Principe de la sélection de Modèle

A partir d'un élément aléatoire $\mathcal{W}$ (vecteur d'observation) de loi inconnue P (de densité f), l'objet de la sélection de modèle est d'approcher P (ou une fonction de P). Le principe est de proposer un modèle $\mathcal{M}$, défini par un ensemble de lois possibles $\mathcal{M} = \{P_\theta, \theta \in \Theta\}$, de supposer que $P = P_\theta$ et de construire P sous la forme $P_{\hat{\theta}(\mathcal{W})}$. Deux approches sont possibles : l'approche paramétrique où Θ est un espace euclidien et l'approche non-paramétrique où Θ est un espace fonctionnel.
Comme P est inconnue, il est difficile de savoir si le modèle considéré est approprié. Dans la mesure où le modèle essayé est inadéquat, l'inférence basée sur les données et le modèle sera mauvaise. Ainsi plus le modèle est restreint et plus le risque de considérer une loi inappropriée est grand. Ce risque est appelé risque d'approximation. D'un autre côté, plus le modèle est riche et plus on a de chance de disposer d'un modèle P_{θ_0} proche de la réalité, mais en contre-partie $P_{\hat{\theta}(\mathcal{W})}$ a alors un risque (en terme d'estimation) d'être loin de P_{θ_0} et donc de P. L'objet de la sélection de modèle est de faire un compromis entre le risque d'approximation et le risque d'estimation. Nous shématisons ce problème de parcimonie au niveau d'un modèle par la figure ci-dessous :



$d\{\cdot;\cdot\}$ représente une mesure d'information qui sera définie ultérieurement et symbolise une distance entre deux distributions. La quantité $d\{P; P_{\theta_0}\}$ est appelé " un terme de biais" et s'annule lorsque la vraie loi appartient au modèle considéré. Ce terme traduit la qualité d'approximation du modèle. La quantité $d\{P_{\widehat{\theta}(\mathcal{W})}; P_0\}$ est appelé "terme de variance". Il est dû à l'estimation faite et est d'autant plus grand que le modèle considéré est grand. Ce terme traduit l'erreur d'estimation. L'art de la sélection de modèles est de définir un modèle qui réalise le meilleur compromis entre le terme de biais et le terme de variance. Un moyen de sélectionner le meilleur des N modèles $\mathcal{M}^1, \mathcal{M}^2, \ldots, \mathcal{M}^N$ est de déterminer les N distances $d\{P; P^k_{\widehat{\theta}(\mathcal{W})}\}$ ($k \in \{1, \ldots, N\}$) représentant l'éloignement des estimateurs $P^k_{\widehat{\theta}(\mathcal{W})}$ à la réalité P. Nous représentons ci-dessous cette pratique où $\left\{\mathcal{M}^k\right\}_{k=1,\ldots,N}$ défini une famille de N modèles. L'objectif est de déterminer le meilleur modèle au sens d'une "distance" que nous présentons dans le paragraphe suivant.



Définition d'une distance

La théorie de l'information [7], et plus particulièrement de l'information de Kullback-Leibler [26], est le principal fondement théorique de la sélection de modèles. L'information de Kullback-Leibler se présente comme l'opposée de l'entropie de Boltzman développée en physique et en thermodynamique. On parle aussi de distance ou encore de divergence de Kullback-Leibler. La motivation du travail de Kullback-Leibler était de fournir une définition rigoureuse de l'information en relation avec la statistique suffisante de Fischer. Cette information entre deux modèles f et g n'est pas exactement une distance car la mesure entre f et g est différente de la mesure entre g et f. Elle est définie entre deux modèles représentés par les densités f et g (supposées continues) par la relation :

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

Cette quantité représente l'information perdue quand g est utilisée pour approcher f. Elle est toujours positive et ne devient nulle que lorsque $f \equiv g$. Akaike [1] se base sur cette information pour définir le AIC (an information criterion) comme critère de sélection dans un cadre paramétrique. Il montre que la log-vraisemblance maximisée ($\mathcal{L}$) est un estimateur biaisé de l'information de Kullback-Leibler. Sous certaines conditions ([25]),

ce biais est égal au nombre de paramètres estimés (K). Ainsi, un estimateur non biaisé de l'espérance de l'information de Kullback-Leibler est $\log \mathcal{L} - K$. En multipliant par -2, le critère d'Akaike est défini par la relation $AIC = -2 \log \mathcal{L} + 2K$ (une construction du AIC est présentée en annexe A).
En définissant par $\left\{ g^k_{\widehat{\theta}(\mathcal{W})} \right\}_{k=1,\ldots,N}$ la famille d'estimateurs correspondant à la famille de modèles $\left\{ \mathcal{M}^k \right\}_{k=1,\ldots,N}$, la sélection de modèles revient à choisir l'estimateur $g^k_{\widehat{\theta}(\mathcal{W})}$ qui minimise l'information de Kullback-Leibler $I\{f, g^k_{\widehat{\theta}(\mathcal{W})}\} \equiv d\{P, P^k_{\widehat{\theta}(\mathcal{W})}\}$. Finalement, notre intérêt réside donc dans la famille d'estimateurs et non plus dans la famille de modèles. Il est évident qu'à chaque modèle, il existe une procédure d'estimation conduisant à un estimateur. En revanche, certaines procédures d'estimations telle que la méthode à noyau, permettent de déterminer un estimateur sans qu'il soit possible de définir un modèle stricto sensu. Dans tous les cas, la difficulté se situe dans l'estimation de l'information $I\{f, g^k_{\widehat{\theta}(\mathcal{W})}\} \equiv d\{P, P^k_{\widehat{\theta}(\mathcal{W})}\}$ puisque la réalité représentée par f ou P est inconnue.
Comme alternative possible à l'information de Kullback-Leibler, d'autres auteurs ([23], [28]) se sont intéressés au MISE (mean integrated square error). Dans un cadre non-paramétrique, cette autre distance a principalement permis la définition de critères afin de sélectionner le paramètre de lissage ([10], [24], [12]).

Plan de la thèse

Le premier chapitre résoud un problème des tests d'hypothèses dans la sélection de modèles. Dans une optique de recherche de l'effet d'une variable explicative, les épidémiologistes usent et abusent de tests afin de trouver le meilleur codage de cette variable. Ils retiennent souvent le test ayant la p_{value} minimale et omettent de corriger le niveau de signification (p_{value}) de ce test. Nous proposons une méthode et un programme permettant de corriger le niveau de signification lorsque plusieurs tests sont effectués. Ce travail est réalisé dans le cadre d'une régression logistique.

Dans le deuxième chapitre, nous proposons un critère d'information permettant de

choisir un estimateur parmi une famille d'estimateurs semi-paramétriques. Le critère proposé est une extension du EIC proposé par Ishiguro, Sakamoto et Kitagawa [21] dans un cadre paramétrique. Il est basé sur l'estimation par bootstrap de l'information de Kullback-Leibler.

Le chapitre 3 présente une application concernant le risque de Mésothéliome pleural lié à une exposition professionnelle à l'amiante. En collaboration avec le laboratoire Santé Travail et Environnement, nous nous sommes intéressés à représenter l'effet de l'amiante sur le risque de mésothéliome. Nous proposons d'utiliser le critère d'information développé au chapitre 2 afin de définir la modélisation la mieux adaptée à ces données. Des estimateurs paramétriques et semi-pamétriques sont essayés. Nous appliquons également dans ce chapitre, la méthode de sélection de modèles de Birgé-Massart [3] pour définir l'estimateur constant par morceaux représentant le mieux l'effet de l'amiante sur le risque de mésothéliome.

Le chapitre 4 présente une extension du critère proposé dans le chapitre 2 dans le cas de données incomplètes. Le critère proposé permet de sélectionner le paramètre de lissage dans l'estimation lisse d'une fonction de risque. Plus généralement, ce critère permet de choisir entre plusieurs modélisations. Des modèles à risque proportionnel ont été comparés à des modèles stratifiés. Une application à une étude épidémiologique concernant le risque de démence chez les personnes âgées illustre le problème de choix de modèles.

Bibliographie

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

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 :716–723, 1974.

[3] L. Birgé and P. Massart. J. Eur. Math. Soc. *Gaussian model selection*, 3 :203–268, 2001.

[4] P. Burman. A comparative study of ordinary cross-validation, ν-hold cross-validation and repeated learning-testing methods. *Biometrika*, 76 :503–514, 1989.

[5] K. P. Burnham and D. R. Anderson. *Model selection and inference : a practical information theoretic approach.* Springer-Verlag, New York, 1998.

[6] J. B. Copas. Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society B*, 45 :311–354, 1983.

[7] T.M. Cover and J.A. Thomas. *Elements of information theory*, page 542. John Wiley and Sons, New York, NY, 1991.

[8] P. Craven and 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.

[9] N. R. Draper. Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B*, 57 :45–97, 1995.

[10] J.D. Fermanian. A new bandwidth selector in hazard estimation. *Nonparametric Statistics*, 10 :137–182, 1999.

[11] S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70 :320–328, 1975.

[12] W. Gonzalez-Manteiga, R. Cao, and J.S. Marron. Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. *Journal of the American Statistical Association*, 91 :1130–1140, 1996.

[13] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London, 1994.

[14] P. J. Green and B. Yandell. Semi-parametric generalized linear models. In *Generalized Linear Models, Lecture Notes in Statistics*, volume 32, pages 44–55. Springer-Verlag, Berlin, 1985.

[15] C. Gu. Cross-validating non-gaussian data. *Journal of Computational and Graphical Statistics*, 1 :169–179, 1992.

[16] W. Hardle. *Applied Nonparametric Regression*. Cambridge University Press, Cambridge, 1990.

[17] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, London, 1990.

[18] J.S.U. Hjorth. *Computer intensive statistical methods : validation, model selection and bootstrap*. Chapman and Hall, London, 1994.

[19] C. M. Hurvich, J.S. Simonoff, and C.L Tsai. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society B*, 60 :271–293, 1998.

[20] C. M. Hurvich and C.L Tsai. Regression and time series model selection in small samples. *Biometrika*, 76 :297–307, 1989.

[21] M. Ishiguro, Y. Sakamoto, and G. Kitagawa. Bootstrapping log likelihood and EIC, an extension of AIC. *Annals of the Institute of Statistical Mathematics*, 49 :411–434, 1997.

[22] P. Joly, D. Commenges, and L. Letenneur. A penalized likelihood approach for arbitrarily censored and truncated data : application to age-specific incidence of dementia. *Biometrics*, 54 :185–194, 1998.

[23] M. C. Jones. The roles of ise and mise in density estimation. *Probability Letters*, 12 :51–56, 1991.

[24] M. C. Jones, J. S. Marron, and Sheather S. J. Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*, 11 :337–38, 1996.

[25] S. Konishi and G. Kitagawa. Generalised information criteria in model selection. *Biometrika*, 83 :875–890, 1996.

[26] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22 :79–86, 1951.

[27] C.L. Mallows. Some comments on c_p. *Technometrics*, 15 :661–675, 1973.

[28] J. S. Marron and M.P. Wand. Exact mean integrated squared error. *Annals of Statistics*, 20 :712–736, 1992.

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

[30] W. Pan. Bootstrapping likelihood for model selection with small samples. *Journal of Computational and Graphical Statistics*, 8 :687–698, 1999.

[31] B. M. Potscher. Effects of model selection on inference. *Econometric Theory*, 7 :163–185, 1991.

[32] J.O. Rawlings. *Applied regression analysis : a research tool.* Wadsworth, Inc., Belmont, CA, 1988.

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

[34] S.L. Sclove. Some aspects of model-selection criteria. In H. Bozdogan, editor, *Engineering and Scientific Applications*, volume 2, pages 37–67, Dordrecht, Netherlands, 1994. Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling : An Informational Approach, Kluwer Academic Publisher.

[35] J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422) :486–494, 1993.

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

[37] B. W. Silverman. Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society B*, 46 :1–52, 1985.

[38] B.W. Silverman. *Density estimation for statistics and data analysis*. Chapman and Hall, London, 1986.

[39] 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.

[40] M. Stone. Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B*, 39 :111–147, 1974.

[41] P.H. Westfall and S.S. Young. *Resampling-based multiple testing : examples and methods for p-value adjustment*. John Wiley and Sons, new york, ny, 1993.

[42] P. Zhang. Model selection via multifold cross-validation. *Annals of Statistics*, 20 :299–313, 1993.

Chapitre 1

Correction du degré de signification après des tests multiples dans une régression logistique

1.1 Introduction

Dans ce chapitre, nous nous intéressons à un problème simple de sélection de modèle qui consiste à choisir la transformation optimale d'une variable quantitative représentant un facteur prognostique ou un facteur de risque.
La transformation la plus usuelle est la transformation dichotomique. Dans ce cas, un seul point de coupure permet de séparer les patients en deux catégories : ceux avec un risque élevé et ceux avec un risque faible. Ces catégories sont souvent utilisées pour faire des recommandations sur les traitements utilisés dans des essais thérapeutiques. En général, afin de choisir la meilleure séparation, plusieurs points de coupure sont testés. La sélection du meilleur point de coupure correspond à la P_{value} minimale associée au test de la variable recodée. Cette approche est nommée la recherche de la P_{value} minimale [1]. La transformation dichotomique permet de représenter un effet seuil de la variable explicative sur la variable dépendante. Souvent, la forme de l'effet est inconnue et d'autres

transformations de la variable explicative sont possibles [3]. De ce fait, le statisticien est amené à utiliser plusieurs transformations afin de choisir celle qui semble la plus favorable. Généralement, les épidémiologistes ne tiennent pas compte de cette procédure d'optimisation et donnent simplement comme résultat final la P_{value} minimale correspondant au test le plus significatif. Nous proposons une méthode et un progamme pour déterminer le niveau de signification pour plusieurs codages d'une variable explicative dans une régression logistique. Ce travail a inspiré une étude similaire afin de résoudre ce problème dans un modèle de Cox [2].

Dans la suite de ce chapitre, nous présentons notre article publié dans "Statistics in Medicine", intitulé "Correction of the P-value after multiple coding of an explanatory variable in logistic regression" [4]

Bibliographie

[1] D. G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher. Dangers of using 'optimal' cutpoints in the evaluation of prognostic factos. *Journal of the National Cancer Institute*, 86 :829–835, 1994.

[2] R. Hashemi and D. Commenges. Correction of the p-value after multiple tests in a cox proportional hazard model. *Lifetime Data Analysis*, 8 :335–348, 2002.

[3] R. Kay and S. Little. 'Transformation of the explanatory variables on the logisic regression model for binary data'. *Biometrika*, 74 :495–501, 1987.

[4] B. Liquet and D. Commenges. Correction of the p-value after multiple coding of an explanatory variable in logistic regression. *Statistics in Medicine*, 20 :2815–2826, 2001.

CORRECTION OF THE P_{value} AFTER MULTIPLE CODING OF AN EXPLANATORY VARIABLE IN LOGISTIC REGRESSION

B. LIQUET[1] and D.COMMENGES

INSERM U330, ISPED

146 rue Léo Saignat

33076 Bordeaux cedex, France

Summary

We propose a method and a program to determine a significance level for a series of codings of an explanatory variable in logistic regression. Dichotomous and Box-Cox transformations are considered. Three methods of correcting the significance level are studied: Bonferroni method, Efron's method which uses the correlation between successive tests, and the exact calculation by numerical integration using all correlations. A simulation study has led to a strategy for the choice and number of the different codings of the variable. This method is illustrated using the data of a study of the relation between cholesterol and dementia.

[1]**E-mail: liquet@isped.u-bordeaux2.fr**

1.2 Introduction

The relationship between an explanatory variable and a dependent variable can be investigated by a statistical model allowing an estimation of the magnitude of this effect. In models such as the linear or logistic regression models, the continuous covariates are often transformed. Binary coding is often used in epidemiology to make interpretation easier or because a threshold effect is suspected. Such is the case, for example, in a study explaining the probability of success of an in-vitro fertilization embryo transfer process where this coding is used for age [1]. The log transformation [2] is also often used, as in a study about cognitive impairment and concentration of minerals in drinking water [3], in which the authors consider the log of aluminium concentration. Other transformations are also used, in particular Box-Cox transformations [4], but the choice of the transformation is often subjective. Hence, to analyse data, the statistician may have to use several transformations and choose the most favourable one. When testing several codings of a variable, there is a problem with the multiplicity of tests done leading to incorrect P_{value} and overestimation of effects [5]. Generally, researchers fail to consider this problem and do not correct the significance level in relation to the number of tests performed.
Numerous methods of correcting the significance level are used to resolve the problem of test multiplicity in clinical trials [6, 7] and in other multiple comparison situations [8, 9]. The Bonferroni method is the simplest and best known approach. Several authors [10, 11, 12, 13] have improved this method to make it more powerful. In some situations, the tests are correlated, and the correlations increase with the number of tests performed. Some authors have considered what happens at the limit when an infinity of tests is performed, an approach which may be relevant when a very large number of contingency tables is considered [14, 15, 16, 17, 18], or in genetic epidemiology when many markers are tested [19]. However, only a limited number of transformations is normally done. Efron [20] has proposed a method which takes into account the correlation structure among successive tests. Nevertheless, if the dependence between the tests is known, the significance level can be obtained by computing a multiple integral.

The goal of the present study is to propose a method and a program to determine the significance level for a series of several codings of an explanatory variable in logistic regression, a model widely used in epidemiology [21] (the problem of correcting the estimation of the effect will not be treated here).

First, we present our motivating example about a study of the relation between cholesterol and dementia [22]. In section 1.4 the Bonferroni method, Efron's method and the exact calculation will be described. Then we study the correlation between the tests in logistic regression. Next the various transformations of an explanatory variable are exposed. Section 1.7 presents simulations which may help to choose a strategy when the shape of the effect is unknown. The selected strategy is then applied to our example. Finally, a description of a program is given in the last section.

1.3 Example

This is a study of the relationship between cholesterol and dementia done by Bonarek *et al.* [22]. This study deals with the analysis of the biological data of a cohort of community dwelling elderly persons, composed of 334 subjects, 37 demented and 297 non-demented. Age, sex, level of education, and wine consumption were considered as adjustment variables. In particular, the influence of HDL-cholesterol (High-Density Lipoprotein) on the risk of dementia was investigated. Bonarek *et al* first considered HDL-cholesterol as a continuous variable; then for easier clinical interpretation, HDL-cholesterol was transformed into a categorical variable with four classes. Finally, as there was no significant difference between the first three quartiles, HDL-cholesterol was split into two categories with a cutpoint at the last quartile. The best P_{value}, 0.007, was obtained in the latter analysis. However this P_{value} did not take into account the numerous transformations performed to determine the best representation of the variable of interest; one may doubt whether there is really a significant association between HDL-cholesterol and dementia.

1.4 Presentation of different methods

Suppose that we take an explanatory variable transformed using K different codings. Each coding (k) corresponds to a statistical test statistic (T_k). Thus we have a sequence of test statistics $T = (T_1, ..., T_K)$ for the same null hypothesis H_0. The T_k have the standard normal distribution. Rejecting H_0 if one of the test T_k is larger than a critical value c is equivalent to rejecting H_0 if $T_{max} > c$, where $T_{max} = max(T_1, ..., T_K)$.
The Bonferroni method has been described by several authors [10, 13] in various applications. It makes it possible to compute an upper bound of the significance level (P_{value}) of the statistic T_{max} :

$$P(T_{max} > t_{max}) \leq \overline{\Phi}(t_{max})K,$$

where

$$\overline{\Phi}(t_{max}) = \int_{t_{max}}^{\infty} \phi(t)dt, \quad \phi(t) = e^{-t^2/2}/(2\pi)^{\frac{1}{2}}.$$

This method is very simple and does not require any assumption of the correlations between the different tests. It can therefore directly be applied to the different possible codings of the explanatory variable. However, it only provides an upper bound of the P_{value}, which may be very conservative if the correlations between tests are high.
Efron [20] has proposed a method which takes into account the correlations between successive tests. When the tests are ordered according to the correlations ($\mathrm{corr}(\mathrm{T_k}, \mathrm{T_{k+1}}) \geq \mathrm{corr}(\mathrm{T_k}, \mathrm{T_{k+h}})$, $h = 1, 2, ..., K-k$ and $k = 1, 2, ..., K$), Efron suggests the following upper bound :

$$P(T_{max} > t_{max}) \leq \overline{\Phi}(t_{max}) + \phi(t_{max}) \sum_{k=1}^{K} \frac{\Phi(t_{max}L_k/2) - \frac{1}{2}}{t_{max}/2}$$

where

$$L_k = arccos(\rho_k), \quad \rho_k = corr(T_{k-1}, T_k).$$

If the tests are well ordered, it is possible with Efron's method to get closer to the exact P_{value}. If they are not, this method is less effective. Moreover, like the Bonferroni method,

Efron's method only yields an upper bound. However, it is possible to compute the P_{value} exactly by writing :

$P_{value} = P(T_{max} > t_{max}) = 1 - P(T_{max} < t_{max}) = 1 - P(T_1 < t_{max}, ..., T_K < t_{max})$

and computing $P(T_1 < t_{max}, ..., T_K < t_{max})$ by numerical integration of the density of the multivariate normal distribution. Several programs have been written to solve the numerical problems set by this multiple integral. Gens' program [23] is used in the present study. What is interesting in this exact computation is its application for all the possible transformations of the explanatory variable. On the other hand, it requires the calculation of the correlations between all the tests, whereas Efron's method only uses the correlation between successive tests.

1.5 Correlation between tests in logistic regression

1.5.1 Score test

Let us consider a logistic model with p explanatory variables where Y_i is a binary dependent variable. We want to test the influence of a variable z_p^i so we consider K transformations of this variable $z_p^i(1), ..., z_p^i(k), ..., z_p^i(K)$. The model for transformation k can be written :

$$\begin{aligned} \text{Logit}(\pi_i) &= \beta_0 + \beta_1 z_1^i + ... + \beta_p z_p^i(k) \\ \pi_i &= P(Y_i = 1|Z^i) = \frac{e^{Z^i\beta + \beta_p z_p^i}}{1 + e^{Z^i\beta + \beta_p z_p^i}} \qquad \imath = 1, ..., n \end{aligned}$$

where n is the number of observations, $\beta = (\beta_0, \beta_1, ..., \beta_{p-1})^T$ and $\mathbf{Z}^i = (1, z_1^i, z_2^i, ...z_{p-1}^i)$. For each transformation k, "$\beta_p = 0$" leads to the same null hypothesis, i.e. the model with explanatory variables $z_1, ..., z_{p-1}$. The score test has the simplest form and is asymptotically equivalent to the Wald test and to the likelihood ratio test. The score test (T_k), under the null hypothesis $\beta_p = 0$, has asymptotically a standard normal distribution. It

is obtained by normalising the score statistic $U_p(k)$:

$$U_p(k) = \frac{\partial \ln L}{\partial \beta_p}(\beta_p = 0) = \sum_{i=1}^{n} z_p^i(k)[Y_i - \hat{\pi}_i] = Z_p^T(k)(Y - \hat{\pi}) \tag{1.1}$$

where L is the likelihood, $\hat{\pi}_i$ is estimated using the maximum likelihood estimator $\hat{\beta}$ of β, under the null hypothesis (i.e. with the model excluding the variable z_p), and $Z_p(k)$, Y and $\hat{\pi}$ are the vectors of entries $z_p^i(k)$, Y_i, $\hat{\pi}_i$ respectively. The variance of the score test is:

$$\text{var } U_p(k) = Z_p^T(k)[(I - H)V]Z_p(k),$$

where $H = VZ[Z^TVZ]^{-1}Z^T$, Z is a $n \times p$ matrix with rows Z^i, $i = 1, .., n$ and V is the diagonal matrix with diagonal terms $v_{ii} = \pi_i(1 - \pi_i)$. This result is known and can be obtained by using the general theory of score tests [24].

1.5.2 Correlation between two tests

Let T_k and T_l be two score test statistics associated to the transformations $z_p(k)$ and $z_p(l)$.

$$T_k = \frac{U_p(k)}{\sqrt{\text{var}[U_p(k)]}} \quad ; \quad T_l = \frac{U_p(l)}{\sqrt{\text{var}[U_p(l)]}}$$

The correlation between the two tests is written: $\rho_{kl} = \text{corr}(T_k, T_l) = \text{cov}(T_k, T_l)$ because T_k and T_l are normalised, (var T_k = var $T_l = 1$).
Neglecting the covariance between the estimators of the variances of $U_p(k)$ and $U_p(l)$, equation (1.1) gives:

$$\rho_{kl} = \text{cov}(T_k, T_l) \simeq \frac{\text{cov}(U_p(k), U_p(l))}{\sqrt{\text{var } U_p(k)}\sqrt{\text{var } U_p(l)}} = \frac{Z_p^T(k)\text{var}[Y - \hat{\pi}]Z_p(l)}{\sqrt{\text{var } U_p(k)}\sqrt{\text{var } U_p(l)}}.$$

Using a Taylor expansion, le Cessie and van Houwelingen [25, 26] showed that:

$$Y - \widehat{\pi} \simeq (I - H)(Y - \pi) \tag{1.2}$$

By using (1.2) and the fact that $\text{var}(Y) = V$, the variance of $Y - \widehat{\pi}$ can be written: $\text{var}(Y - \widehat{\pi}) \simeq (I - H)V$ (because $I - H$ is a projection matrix), so that

$$\rho_{kl} \simeq \frac{Z_p^T(k)(I - H)VZ_p(l)}{\sqrt{\text{var } U_p(k)}\sqrt{\text{var } U_p(l)}}.$$

which is consistent with the formula for var $U_p(k)$, putting k=l ($\rho_{kk} = 1$).

1.6 Transformations

Two kinds of transformations were considered: dichotomous and Box-Cox transformations [27].

1.6.1 Dichotomous transformations

Dichotomous transformations are defined as:

$$z_p(k) = \begin{cases} 0 & \text{if } z_p \leq c_k \\ 1 & \text{if } z_p > c_k \end{cases}$$

To obtain the best transformation, several cutpoints may be tested. When no epidemiological references are available, the most usual strategy consists in choosing the median for a dichotomous transformation and taking the first tercile, then the second, as cutpoints for two dichotomous transformations and so on. This strategy is summarised in table 1.1.

Table 1.1: Strategy for the dichotomous transformations: values of c_k according to the number of transformations.

number of transformations	c_1	c_2	c_3	c_4	...	c_9
1	median					
2	1^{st}tercile	2^{nd}tercile				
3	1^{st}quartile	2^{nd}quartile	3^{rd}quartile			
4	1^{st}quintile	2^{nd}quintile	3^{rd}quintile	4^{th}quintile		
⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	1^{st}decile	2^{nd}decile	3^{rd}decile	4^{th}decile	...	9^{th}decile

Dichotomous transformations can be applied to continuous variables or to ordered categorical variables. While for a continuous variable the number of possible cutpoints is infinite, for an ordered categorical variable with J levels, only $J-1$ distinct dichotomous variables are possible.

1.6.2 Box-Cox transformations

The family of Box-Cox transformations is defined by:

$$z_p(k) = \begin{cases} \lambda_k^{-1}(z_p^{\lambda_k} - 1) & \text{if } \lambda_k > 0 \\ \log z_p & \text{if } \lambda_k = 0 \end{cases}$$

In particular, $\lambda_k = 1$ implies no transformation, $\lambda_k = 0$ gives a log transformation and $\lambda_k = 0.5$ a square root. There is no obvious strategy here as for dichotomous transformations. It seems natural to try the crude variable ($\lambda_1 = 1$) and since the log transformation is often interesting, we propose $\lambda_1 = 1$ and $\lambda_2 = 0$ when two transformations are tried. Further interesting transformations include the square and the square root. Finally, the power 3/2 may be tried. The proposed strategy for this family of transformations is summarised in table 1.2.

Table 1.2: Strategy for the Box-Cox transformations: values of λ according to the number of transformations

number of transformations	λ_1	λ_2	λ_3	λ_4	λ_5
1	1				
2	1	0			
3	1	0	2		
4	1	0	2	0.5	
5	1	0	2	0.5	1.5

1.7 Simulations

Simulations were carried out with a logistic model consisting of two explanatory variables (z_1 adjustment variable and z_2 tested variable):

$$Logit(\pi_i) = \beta_0 + \beta_1 z_1^i + \beta_2 z_2^i$$

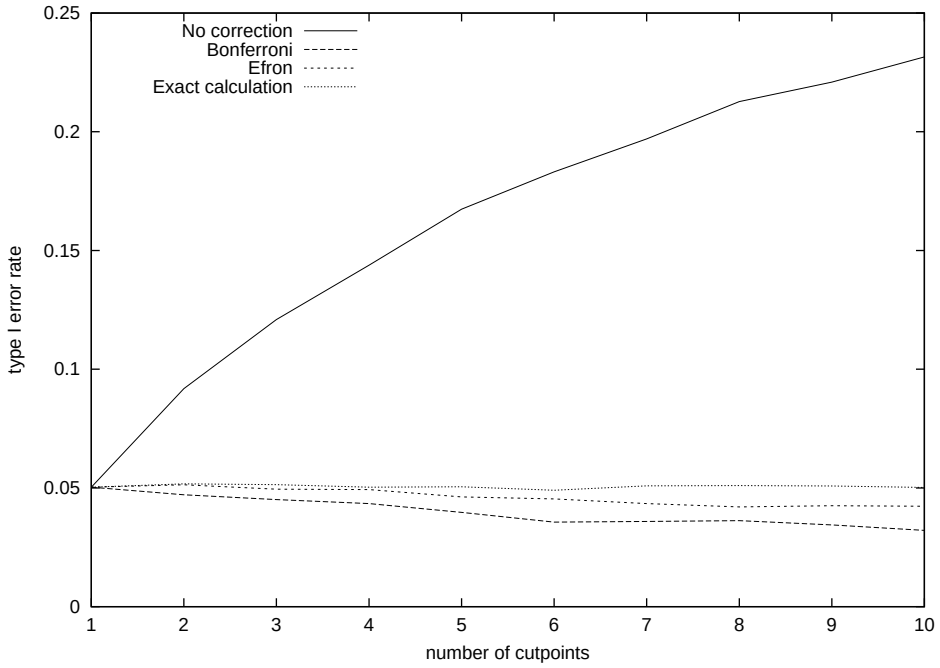
z_1 and z_2 were generated according to a uniform distribution ($z_1 \sim U[a,b]$, $z_2 \sim U[a,b]$). In the simulations, we arbitrarily chose a=0 and b=4. The sample size was set to be 100. We used 10,000 replications for each simulation.

1.7.1 Study of type I error rate

For a replication, the rejection criterion of H_0 ($\beta_2 = 0$) was a P_{value} less than 0.05. Thus, for a simulation (10,000 replications), the empirical type I error rate was the proportion of the number of times the P_{value} was less than 0.05. Figure 1.1 shows the type I error rate for dichotomous transformations.

The "no correction" curve showed the type I error rate without taking into account the multiplicity of tests. This error rate increased with the number of cutpoints tried. The error rate calculated by the Bonferroni method decreased with the number of cutpoints.

Figure 1.1: Type I error rate according to the different methods for dichotomous transformations ($\beta_0 = 1$, $\beta_1 = 2$)



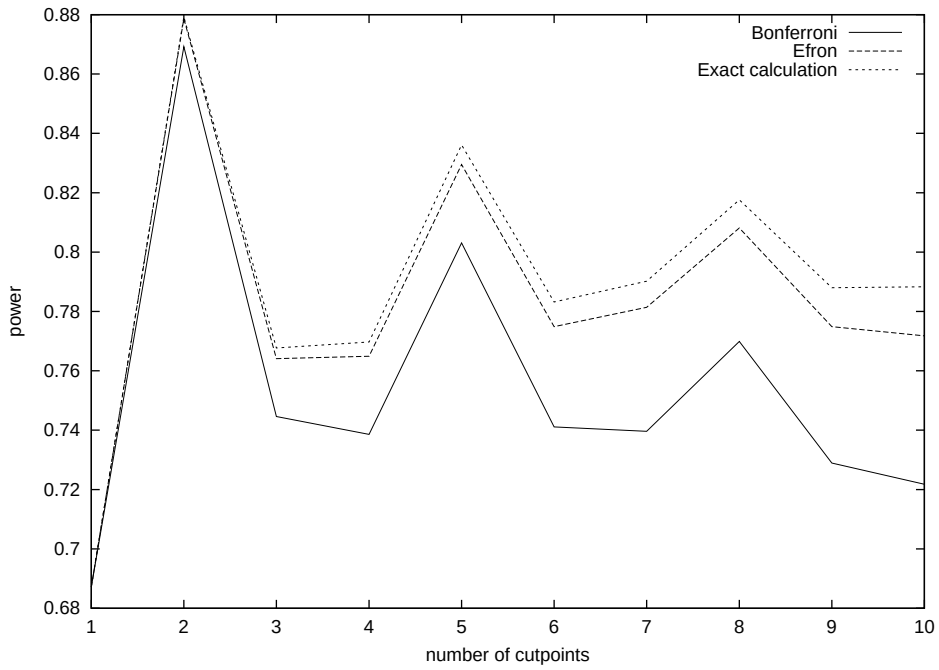
This correction was therefore too conservative whereas Efron's method, which was slightly conservative as expected, approached the exact calculation. This exact calculation gave the type I error rate closest to the nominal value.

1.7.2 Power

The first simulation represented the situation in which the epidemiologist had some information on the shape of the effect of the explanatory variable. To begin with, we studied the power for a threshold effect model with a cutpoint value at the first tercile ("threshold effect model"). Figure 1.2 gives the power as a function of the numbers of cutpoints tried.

The power of the exact calculation was the highest whatever the number of cutpoints. It was maximal with two cutpoints. This result was expected since using two cutpoints amounts to testing the dichotomized variable in its first and second tercile. Power increased again when trying five cutpoints because the value of the second test corresponds

Figure 1.2: Power for a "threshold effect model" at the first tercile ($\beta_0 = -2.5,\ \beta_1 = 1,\ \beta_2 = 1.5$)



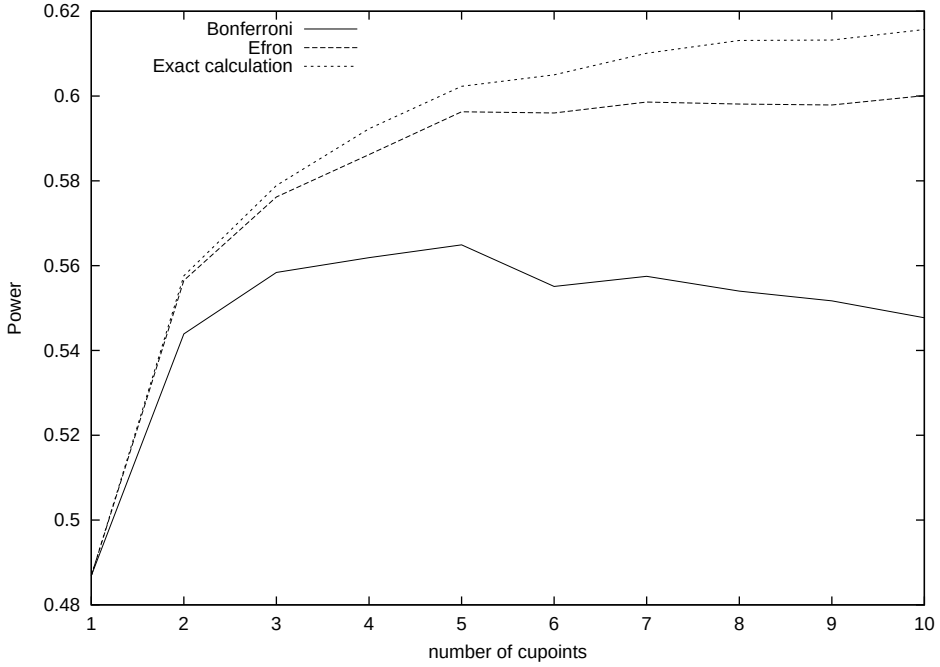
to the first tercile. This simulation showed that the exact calculation provides more power than the Bonferroni and Efron methods. However, this situation was theoretical. In general, a threshold effect may be assumed but the cutpoint is unknown and is not expected to fall on a simple quantile (median, tercile, quartile ...). It was possible to represent the epidemilogist's uncertainty by considering that the cutpoint had the same probability at any quantile. This situation was modelled as follows :

$$Logit(\pi) = \beta_0 + \beta_1 z_1 + \beta_2 z_2(c)$$

where the dichotomous variable $z_2(c)$ was obtained from the continuous variable z_2 (with a distribution F) by selecting a cutpoint c with the same distribution F as z_2. For instance, if $z_2 \sim U[0,4]$ then $c \sim U[0,4]$. Figure 1.3 shows the results of this simulation.

The exact calculation gave the highest power. Regarding this simulation, when a threshold effect is supposed, it seems to be sufficient to perform 6 or 7 cutpoints (power=

Figure 1.3: Power for a random "threshold effect model" ($\beta_0 = -2.5,\ \beta_1 = 1,\ \beta_2 = 1.5$)



0.603 at 6 cutpoints), since beyond 6 cutpoints little additional power was obtained (power= 0.605 at 10 cutpoints).

Moreover, we considered the case where the true model involved a Box-Cox transformation of z_p ("Box-cox effect model"). However, an epidemiologist will not know which particular transformation should be used. To illustrate this situation, we simulated a model as follows:

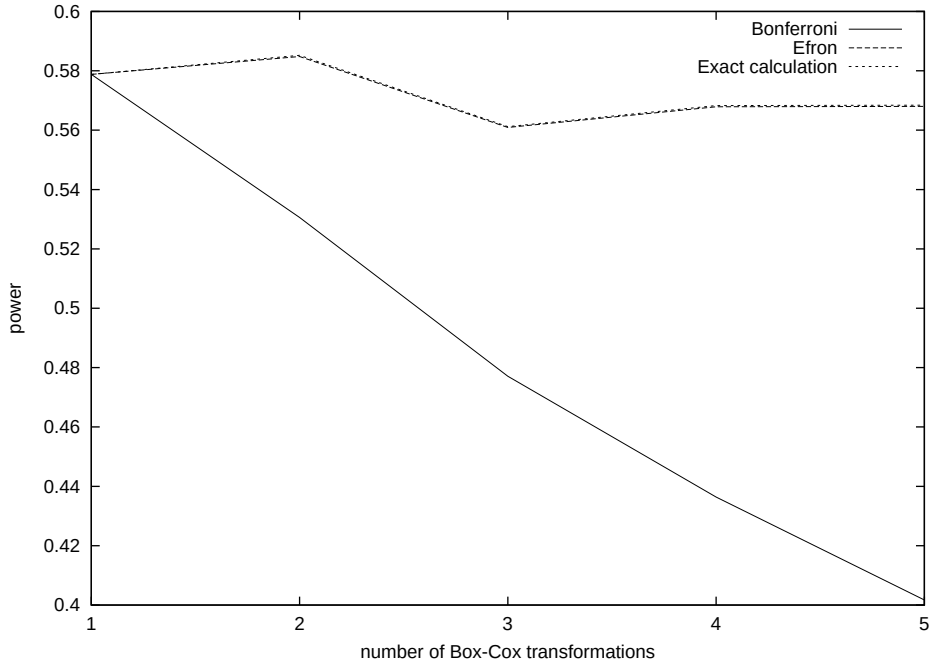
$$Logit(\pi) = \beta_0 + \beta_1 z_1 + \beta_2 \frac{BC_\lambda(z_2)}{(\text{var } BC_\lambda(z_2))^{\frac{1}{2}}}$$

where $BC_\lambda(z_2)$ was a Box-Cox transformation of z_2 with a parameter λ. We had to standardise the transformed variable to get a constant average effect (effect of an increase of one standard deviation in the explanatory variable), because the Box-Cox transformation modifies the variance of the variable. Without this standardisation, for some λ_k, the effect would be so great that the null hypothesis would be rejected whatever the test used; or so weak that it would never be rejected. In this simulation, the distribution of the continuous

variable z_2 may influence the results. We chose z_2 as having a log-normal distribution (with $\mu = e$, $\sigma^2 = e(e-1)$) because many variables of interest in epidemiology (e.g. concentrations of molecules or cells in blood) have a distribution close to the log-normal. For each repetition, λ was generated according to a uniform distribution ($\lambda \sim U[0,2]$). We restricted λ to the interval [0,2] because $\lambda > 2$ would give a very large (frequently too large) effect to extreme values. In epidemiology, it is more usual to reduce the effect of extreme values, as was the case in a study on AIDS [28] where the log of the viral load was used.

Figure 1.4 shows the power calculated for different transformations according to the strategy presented in paragraph 4.2. There was a dramatic loss of power for the Bonferroni

Figure 1.4: Power for a "Box-Cox effect model" ($\beta_0 = -2.5,\ \beta_1 = 3,\ \beta_2 = 1$)



method whereas Efron's method gave nearly the same result as the exact method. Power was approximately stable by exact calculation. It was maximal (power=0.585) with two transformations corresponding to testing the original variable and the log of the variable. Yet we believed it might be interesting to use up to 5 transformations since the loss of

power with more than two transformations was small (power=0.568 at 5 transformations), so this might give information on the shape of the effect.

Frequently, the shape of the effect of the variable is completely unknown, and it may be assumed to be either a threshold or Box-Cox effect. To illustrate this situation, we simulated models which were equally as likely to be of threshold or Box-Cox effect. The power only for the exact method was calculated for the different transformations: the Bonferroni bound is too conservative, while Efron's method is not well suited here because there is no obvious order between the tests. We set the maximum number of dichotomous transformations at 7 which yields more usual cutpoints (median, quartiles) than those obtained with 6 transformations. The maximum number of Box-Cox transformations was set at 5 which made it possible to include the most usual transformations. Table 1.3 presents the results.

Table 1.3: Power for the exact calculation for a threshold effect or Box-Cox effect model ($\beta_0 = -2.5,\ \beta_1 = 1.5,\ \beta_2 = 1.5$)

Number of	Number of dichotomous transformations							
Box-Cox transformations	0	1	2	3	4	5	6	7
0		0.600	0.640	0.667	0.673	0.680	0.686	0.686
1	0.635	0.673	0.680	0.690	0.691	0.691	0.693	0.694
2	0.690	0.693	0.691	0.695	0.695	0.694	0.696	0.696
3	0.671	0.678	0.681	0.685	0.687	0.688	0.691	0.692
4	0.675	0.681	0.684	0.686	0.689	0.689	0.692	0.693
5	0.675	0.681	0.684	0.686	0.689	0.690	0.692	0.693

The power ranged from 0.600 and 0.696. With only one dichotomous or Box-Cox transformation, the power was the lowest. A minimum of 3 dichotomous and one Box-Cox transformations is needed to obtain a good power. Nevertheless, with no epidemiological assumption, as the loss of power is minimal, numerous transformations may be tried. It

is interesting to test up to 5 Box-Cox transformations and 7 cutpoints, because this may give an idea about the shape of the effect of the variable without losing power.

1.8 Application

The proposed strategy was applied to the example described in section 1.3. Our method does not make it possible to compute the corrected P_{value} associated to the categorical variable with more than two categories. However, we applied our strategy to this problem; since no transformation was initially obvious, we tried 7 dichotomous and 5 Box-Cox transformations. The best transformation appeared to be the dichotomous transformation of HDL-cholesterol with a cutpoint at the last quartile, as already found by Bonarek *et al.* The P_{value} accurately calculated was 0.018. Bonferroni correction for the 12 transformations gave a P_{value} equal to 0.087, thus not significant at the usual 0.05 level. On the other hand, the proposed strategy with the exact calculation of the P_{value} gave a result which was still significant and more realistic than the uncorrected P_{value}.

1.9 Program

We propose a program to calculate exactly the P_{value} for a series of several codings of an explanatory variable in a logistic regression model with several adjusting variables. The user can choose between dichotomous or Box-Cox transformations or both. When the user has chosen the number of Box-Cox and dichotomous transformations, two strategies are possible: the program uses the strategy described in section 1.5 to select the transformations or the user chooses his own strategy, indicating the different parameters : values of cutpoints c_k for the dichotomous transformations and values of the λ_k for Box-Cox transformations. The program provides the results of the different score tests as well as the exact P_{value} associated to the best test. The program called CPMTL is written in Fortran 77 and is available on Internet: http://www.isped.u-bordeaux2.fr/ISPED/RECHERCHE/BIOSTATS/FR-BIOSTATS-Accueil.htm

ACKNOWLEDGEMENT:

We thank Luc Letenneur for making the data of the cholesterol study available.

Bibliography

1. Commenges-Ducos M., Tricaud S., Papaxanthos-Roche A., Dallay D., Horovitz J., and Commenges D. 'Modelling of the probability of success of the stages of in-vitro fertilization and embryo transfer: stimulation, fertilization and implantation'. *Human Reproduction*, **14**(3):78–83, (1998).

2. Keene O.N. 'The log transformation is special'. *Statistics in Medicine*, **14**(8):811–819, (1995).

3. Jacqmin-Gadda H., Commenges D., Letenneur L., Barberger-Gateau P., and Dartigues JF. 'Components of Drinking Water and Risk of Cognitive Impairment in the Elderly'. *Am J Epidemiol*, **139**(1):48–57, (1994).

4. Kay R. and Little S. 'Transformation of the explanatory variables on the logisic regression model for binary data'. *Biometrika*, **74**:495–501, (1987).

5. Harrell F.E., Lee K., J.r, and Mark D.B. 'Tutorial in biostatistics. Multivariable prognostic models: issues in developing, evaluating assumptions and adequacy and measuring and reducing errors. *Statistics in Medicine*, **15**:361–387, (1996).

6. Sankoh A.J., Huque M.F., and Dubey S.D. 'Some comments on frequently used multiple endpoint adjustment methods in clinical trials'. *Statistics in Medicine*, **16**:2529–2542, (1997).

7. Pocock S.J. 'Interim Analyses for Randomized Clinical Trials: The Group Sequential Approach'. *Biometrics*, **38**:153–162, (1982).

8. Hochberg Y. and Tamhane A. *'Multiple Comparison Procedures'*. New York:Wiley, 1987.

9. Miller R. *Simultaneous Statistical Inference*. New York:Springer-Verlag, 1981.

10. Simes R.J. 'An Improved Bonferroni procedure for multiple tests of significance'. *Biometrika*, **73**:751–754, (1998).

11. Worsley K.J. 'An Improved Bonferroni Inequality and Applications '. *Biometrika*, **69**:297–302, (1982).

12. Hochberg Y. 'A sharper Bonferroni procedure for multiple test procedure '. *Biometrika*, **73**:751–754, (1988).

13. Sarkar S.K. and Chang C.K. 'The Simes Method for Multiple Hypothesis Testing With Positively Dependent Test Statistics'. *Journal of the American Statistical Association*, **92**:1601–1608, (1997).

14. Mazundar M. and Glassman J.R. 'Tutorial in biostatistics. Categorising a prognostic variable: review of methods, code for easy implementation and applications to decision-making about cancer treatments. *Statistics in Medicine*, **19**:113–132, (2000).

15. Miller R. and Siegmund D. 'Maximally Selected Chi Square'. *Biometrics*, **38**:1011–1016, (1982).

16. Koziol J.A. 'On Maximally Selected Chi-Square Statistics '. *Biometrics*, **47**:1557–1561, (1991).

17. Lausen B. and Schumacher M. 'Maximally Selected Rank Statistics '. *Biometrics*, **48**:73–85, (1992).

18. Lausen B. and Schumacher M. 'Evaluating the effect of optimized cutoff values in the assessment of prognostic factors'. *Computational Statistics Data Analysis*, **21**:307–326, (1996).

19. Lander E. and Kruglyak L. 'Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results'. *Nature Genetics*, **11**:241–247, (1995).

20. Efron B. 'The length heuristic for simultaneous hypothesis tests'. *Biometrika*, **84**(1):143–57, (1997).

21. Hosmer D.W. and Lemeshow S. *Applied Logistic Regression*. Wiley, New York, 1989.

22. Bonarek M., Barberger-Gateau P., Letenneur L., Deschamps V., Iron A., and Dubroca B. 'Relationships between cholesterol, apoplipoprotein E polymorphism and dementia: A cross-sectional analysis from the PAQUID Study'. *Neuroepidemiology*, 1(47):141–149, 2000.

23. Genz A. 'Numerical Computation of Multivariate Normal Probabilities'. *J. of Comutational and Graphical Stat*, **47**(1):141–149, (1992).

24. Cox D.R. and Hinkley D.V. *Theoretical Statistics*, chapter 9, page 324. Chapman and Hall, 1979.

25. Le Cessie S. and Van Houwelingen J.C. 'Testing the fit of a regression model via score tests in random effects models'. *Biometrics*, **51**(2):600–614, (1995).

26. Le Cessie S. and Van Houwelingen J.C. 'A goodness of test for binary regression models, based on smoothing methods'. *Biometrika*, **47**(4):1267–1282, (1991).

27. Guererro V.M. and Johnson R.A. 'Use of the Box-Cox transformation with binary response models'. *Biometrika*, **69**:309–314, (1982).

28. Volberding P. 'HIV quantification : clinical applications'. *Lancet*, **347**:71–73, (1996).

Chapitre 2

Choix par bootstrap d'estimateurs semi-paramétriques

2.1 Introduction

La relation entre une variable dépendante Y et une variable explicative quantitative X se présente généralement sous la forme d'une courbe ($Y = f(X)$ avec f appartenant à une certaine classe de fonction). Deux approches permettent d'estimer la forme de cette courbe f. La première dite "paramétrique" consiste à modéliser f en supposant qu'elle appartient à un modèle $\mathcal{M}$, défini par un nombre fini de paramètres. Le principe du maximum de vraisemblance définit alors un estimateur pour le modèle $\mathcal{M}$. Plusieurs modèles peuvent alors être envisagés et une sélection est faite pour déterminer le meilleur. Le critère AIC, qui est une approximation de Kullback-Leibler, est le critère le plus utilisé pour choisir entre les modélisations paramétriques. Le EIC, une estimation par boostrap de l'information de Kullback-Leibler, peut aussi être utilisé, pour différencier entre les modèles paramétriques. L'autre approche, dite "non paramétrique", consiste à contrôler f par un terme de régularité. L'estimation de f dépend alors d'un paramètre de lissage. Dans des problèmes simples de régression, des versions du AIC sont disponibles pour choisir le paramètre de lissage. Dans un cadre plus général, nous proposons d'étendre et

d'utiliser le EIC afin de choisir entre plusieurs estimateurs parmi une famille d'estimateurs non-(ou semi) paramétriques et une famille d'estimateurs paramétriques.

Dans la suite de ce chapitre, nous présentons notre article (sous presse) dans "Biometrics", intitulé "Bootstrap choice of estimators in parametric and semi-parametric families : an extension of EIC", 2002.

Bootstrap choice of estimators in parametric and semi-parametric families : an extension of EIC

B. LIQUET, C. SAKAROVITCH and D. COMMENGES

INSERM U330, ISPED

146 rue Léo Saignat

33076 Bordeaux cedex, France

Summary

Ishiguro, Sakamoto and Kitagawa (1997) proposed EIC as an extension of Akaike criterion (AIC) ; the idea leading to EIC is to correct the bias of the log-likelihood, considered as an estimator of the Kullback-Leibler information, using bootstrap. We develop this criterion for its use in multivariate non-parametric estimation using a rigourous formalism and argue that it can be used for choosing among parametric and non-parametric estimators. A simulation study based on a regression model shows that EIC is better than its competitors although likelihood cross-validation performs nearly as well except for small sample size. Its use is illustrated by estimating the mean evolution of viral RNA levels in a group of infants infected by HIV.

KEY WORDS : bootstrap, Kullback-Leibler information, regression, semi-parametric, smoothing

2.2 Introduction

Selection of a model in a parametric family has been most often done using AIC (Akaike, 1974) which can be derived as an approximation of the Kullback-Leibler information (Kullback and Leibler, 1951). That is, the selected model has the lowest estimated distance (as defined by Kullback-Leibler information) from the true model. A better approximation, called AICc, has been derived by Sugiura (1978) and Hurvich and Tsai (1989). In deriving these criteria, it is assumed that the parameters are estimated by maximum likelihood and that the true density is in the family of parametric models considered. Ishiguro, Sakamoto and Kitagawa (197) and Konishi and Kitagawa (1996) proposed a bootstrap approach for estimating the Kullback-Leibler information and they denoted the resulting criterion EIC (extended information criterion). Shibata (1997) proved that the EIC is asymptotically equivalent to AIC. The EIC was applied to the variable selection problem and seemed to be better than other criteria for this problem (Ishiguro et al., 1997 ; Pan, 1999).

In this paper we extend the EIC to any parametric or semi-parametric family of estimators. This is presented in section 2.3 where we consider a general multivariate regression problem. We consider the problem of semi-parametric regression and show that the EIC can be used to choose an estimator among a composite set containing families of non- (or semi-) parametric estimators and families of parametric estimators ; we will in particular consider a family of penalised likelihood estimators. The EIC is compared to other criteria available for this problem in a simulation study presented in section 2.4. In section 2.5, the EIC criterion is applied to the problem of estimating the mean HIV RNA level of infected children as a function of age. This study bears on 17 infants infected early during the peripartum period (but not inutero) and for whom 4 measurments of the HIV level were made between birth and 9 months old.

2.3 General Theory

Let $\mathcal{W} = (W_1, \ldots, W_n)$ with $W_i = (Y_i, X_i), (Y_i \in R^p, X_i \in R^q)$ be a sample of independent, identically distributed random variables with common distribution $F_W(\cdot,\cdot) = F_{Y,X}(\cdot,\cdot)$; and density $f_{Y,X}(\cdot|\cdot)$. In the context of regression, we are interested in the conditional distribution of Y given X specified by the density $f_{Y|X}(\cdot|\cdot)$. Consider $\widehat{g}^k_{Y|X}(\cdot|\cdot;\mathcal{W})$ an estimator of $f_{Y|X}(\cdot|\cdot)$ based on $\mathcal{W}$. The parameter k could represent the smoothing parameter in a nonparametric approach or the dimension of a parametric space.
The Kullback-Leibler divergence between $f_{Y|X}(\cdot|X)$ and $\widehat{g}^k_{Y|X}(\cdot|X;\mathcal{W})$

$$\begin{aligned} \mathrm{I}\left\{f_{Y|X}(\cdot|X); \widehat{g}^k_{Y|X}(\cdot|X;\mathcal{W})\right\} &= \int f_{Y|X}(y|X)\log f_{Y|X}(y|X)dy \\ &\quad - \int f_{Y|X}(y|X)\log \widehat{g}^k_{Y|X}(y|X;\mathcal{W})dy \end{aligned}$$

measures how close $\widehat{g}^k_{Y|X}(\cdot|X;\mathcal{W})$ is to the density $f_{Y|X}(\cdot|X)$ for a given X. For comparing different values of k, only the second term is relevant and we denote :

$$\mathrm{KL_k}(X,\mathcal{W}) = \mathrm{E}\left\{\log \widehat{g}^k_{Y|X}(Y|X;\mathcal{W})|X,\mathcal{W}\right\}$$

where $(Y,X) = W$ is a "future" observation and $W \sim F_{Y,X}$, independently of the sample $\mathcal{W}$. Because we wish that our criterion does not depend on X, we define it as

$$\mathrm{EKL_k}(\mathcal{W}) = \mathrm{E}\left\{\mathrm{KL_k}(X,\mathcal{W})|\mathcal{W}\right\} = \mathrm{E}\left\{\log \widehat{g}^k_{Y|X}(Y|X;\mathcal{W})|\mathcal{W}\right\}. \tag{2.1}$$

A natural estimator is found by replacing $F_{Y,X}$ with its empirical distribution $\widehat{F}_{Y,X}(\cdot,\cdot)$:

$$\widehat{\mathrm{EKL}}_k(\mathcal{W}) = \frac{1}{n}\sum_{i=1}^{n}\log \widehat{g}^k_{Y|X}(Y_i|X_i;\mathcal{W}). \tag{2.2}$$

(that we may recognize as the log-likelihood of $\mathcal{W}$ for a conditional probability specified by $\widehat{g}^k_{Y|X}(\cdot|.;\mathcal{W})$)
However this generally overestimates $\mathrm{EKL_k}(\mathcal{W})$ (Bozdogan, 1987). We propose to estimate

the bias $b = \mathrm{E}\left\{\widehat{\mathrm{EKL}}_k(\mathcal{W}) - \mathrm{EKL_k}(\mathcal{W})\right\}$ by the bootstrap. With $\mathcal{W}^* = (W_1^*, \ldots, W_n^*)$ and $W_j^* \sim \widehat{F}_{Y,X}$, we find

$$\widehat{b}(\mathcal{W}) \quad = \quad \mathrm{E}\left\{\widehat{\mathrm{EKL}}_k(\mathcal{W}^*) - \mathrm{EKL_k}(\mathcal{W}^*)|\mathcal{W}\right\}.$$

Replacing $\mathcal{W}$ by $\mathcal{W}^*$ in (2.1) and (2.2), we find :

$$\widehat{b}(\mathcal{W}) \quad = \quad \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log \widehat{g}^k_{Y|X}(Y_i^*|X_i^*;\mathcal{W}^*) - \mathrm{E}\left\{\log \widehat{g}^k_{Y|X}(Y^*|X^*;\mathcal{W}^*)|\mathcal{W}^*,\mathcal{W}\right\}\middle|\mathcal{W}\right]$$

where (Y^*, X^*) is conditionally independent of $\mathcal{W}^*$ given $\mathcal{W}$, and because (Y^*, X^*) in the second term has the distribution $\widehat{F}_{Y,X}$, we have :

$$\widehat{b}(\mathcal{W}) \quad = \quad \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log \widehat{g}^k_{Y|X}(Y_i^*|X_i^*;\mathcal{W}^*) - \frac{1}{n}\sum_{i=1}^{n}\left\{\log \widehat{g}^k_{Y|X}(Y_i|X_i;\mathcal{W}^*)\right\}\middle|\mathcal{W}\right].$$

The expectation conditional on $\mathcal{W}$ can be approximated by a mean of B evaluations with $\mathcal{W}^j$ taken at random from the distribution of $\mathcal{W}^*$:

$$\widehat{b}(\mathcal{W}) \simeq \frac{1}{B}\sum_{j=1}^{B}\left\{\frac{1}{n}\sum_{i=1}^{n}\log \widehat{g}^k_{Y|X}(Y_i^j|X_i^j;\mathcal{W}^j) - \frac{1}{n}\sum_{i=1}^{n}\log \widehat{g}^k_{Y|X}(Y_i|X_i;\mathcal{W}^j)\right\}.$$

Finally our criterion is :

$$\mathrm{EIC} = \frac{1}{n}\sum_{i=1}^{n}\log \widehat{g}^k_{Y|X}(Y_i|X_i;\mathcal{W}) - \widehat{b}(\mathcal{W}).$$

We will choose $\widehat{g}^k_{Y|X}(\cdot|\cdot;\mathcal{W})$ which maximizes EIC.

In practice, we take a bootstrap sample from $\mathcal{W}$. For each bootsrap sample the estimate of the density function is obtained. We then compute the difference between the log-likelihood obtained by this estimate for the bootstrap sample and the log-likelihood obtained by the same estimate for the original sample. This procedure is repeated B times and the average of the differences gives the bias. The EIC itself is computed by removing

the estimated bias from the maximum log-likelihood.

2.4 Simulation

We have done a simulation study to compare different criteria for choosing an estimor of a regression.

2.4.1 Model for the simulation

We observe $\mathcal{W} = (W_1, \ldots, W_n)$ with $W_i = (Y_i, X_i), (Y_i \in R, X_i \in R)$, a sample of independent, identically distributed random variables. We specify $f_{Y|X}(\cdot|\cdot)$ by the model

$$Y_i = h(X_i) + \epsilon_i$$

where $h(\cdot)$ is an unknown function and ϵ_i are independent and normally distributed with expectation 0 and variance σ^2.

2.4.2 Estimators

Several estimators are possible for $f_{Y|X}(\cdot|\cdot)$. We will use a set containing a family of parametric estimators and a family of semi-parametric estimators. The parametric family will be the polynomial estimator $\widehat{g}^k_{Y|X}(\cdot|x; \mathcal{W})$ defined as the density of $\mathcal{N}(\hat{\beta}_0 + \hat{\beta}_1 x + \ldots + \hat{\beta}_{k-1} x^{k-1}; \hat{\sigma}^2)$ where $(\hat{\beta}_0, \ldots, \hat{\beta}_{k-1}, \hat{\sigma}^2)$ maximize the log-likelihood (conditional on $X_i, i = 1, \ldots, n$)

$$-\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_i^2 - \ldots - \beta_{k-1} X_i^{k-1})^2 \qquad k < n-1.$$

The family of semi-parametric estimators will be $\widehat{g}^k_{Y|X}(\cdot|x;\mathcal{W})$ defined as the density of $\mathcal{N}(\hat{h}(x);\hat{\sigma}^2)$ where $\hat{h}$ and $\hat{\sigma}^2$ maximize the penalized log-likelihood (Silverman, 1985) :

$$-\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\{Y_i - h(X_i)\}^2 - k\int h''(u)^2 du.$$

Here k represents the smoothing parameter. We call this family semi-parametric because we do not make any assumption on $h(\cdot)$ but we make the normality asumption for ϵ_i.

2.4.3 Selection criteria

For families of parametric models AIC, AICc and BIC (Schwarz, 1978) are defined as :

$$\begin{aligned}
\text{AIC} &= -2\sum_{i=1}^{n}\log\widehat{g}^k_{Y|X}(Y_i|X_i;\mathcal{W}) + 2k \\
\text{AICc} &= -2\sum_{i=1}^{n}\log\widehat{g}^k_{Y|X}(Y_i|X_i;\mathcal{W}) + 2k\frac{n}{n-k-1} \\
\text{BIC} &= -2\sum_{i=1}^{n}\log\widehat{g}^k_{Y|X}(Y_i|X_i;\mathcal{W}) + k\log n
\end{aligned}$$

The model chosen is that which minimizes the criterion used. The AIC and the AICc criteria have been adapted (Hurvich, Simonoff and Tsai, 1998) for non-parametric estimators which are linear in the sense that $\hat{\mathbf{Y}} = H\mathbf{Y}$, where $\mathbf{Y} = (Y_1,\ldots,Y_n)'$ and $\hat{\mathbf{Y}}$ is the expectation of $\mathbf{Y}$ with respect to the probability defined by the estimator $\widehat{g}^k_{Y|X}(\cdot|x;\mathcal{W})$; H is called the smoother matrix and the trace of H ($\text{tr}(H)$) can be interpreted as the effective number of parameters. Thus, replacing k by $\text{tr}(H)$, the criteria become :

$$\begin{aligned}
\text{AIC} &= -2\sum_{i=1}^{n}\log\widehat{g}^k_{Y|X}(Y_i|X_i;\mathcal{W}) + 2\text{tr}(H) \\
\text{AICc} &= -2\sum_{i=1}^{n}\log\widehat{g}^k_{Y|X}(Y_i|X_i;\mathcal{W}) + 2\text{tr}(H)\frac{n}{n-\text{tr}(H)-1}
\end{aligned}$$

We extended in the same spirit the BIC for non-parametric estimators :

$$\text{BIC} \quad = \quad -2\sum_{i=1}^{n} \log \widehat{g}^{k}_{Y|X}(Y_i|X_i;\mathcal{W}) + \text{tr}(H)\log n$$

The likelihood cross-validation is a general criterion which can also be used in this problem. It is defined as (O'Sullivan, 1988) :

$$\text{LCV} = \sum_{i=1}^{n} \log \widehat{g}^{k}_{Y|X}(Y_i|X_i;\mathcal{W}^{-i})$$

where $\mathcal{W}^{-i} = (\mathcal{W}_1, \ldots, \mathcal{W}_{i-1}, \mathcal{W}_{i+1}, \ldots, \mathcal{W}_n)$. The LCV can be used to choose an estimator among a family containing parametric and non-parametric subfamilies. The cross-validation (CV) is nearly equivalent to the LCV for models with normal noise :

$$\text{CV} = \sum_{i=1}^{n} (Y_i - \widehat{Y}^{-i})^2$$

where $\widehat{Y}^{-i}$ is the estimate of Y_i by the model estimated with $\mathcal{W}^{-i}$. In the case of semi-parametric estimator, the CV can be approximated by the GCV (generalized cross-validation) which involves less computation (Green and Silverman, 1994)) :

$$\text{GCV} \quad = \quad -2\sum_{i=1}^{n} \log \widehat{g}^{k}_{Y|X}(Y_i|X_i;\mathcal{W}) - 2n\log\left\{1 - \frac{tr(H)}{n}\right\}$$

Similary we define the GCV for parametric estimator by replacing $\text{tr}(H)$ by the number of parameters (k).

2.4.4 Results

We performed the simulation with the following parameters :

- sample size n=30, 50 and 100 ;
- pattern of predictor values : a random uniform design on [0,1] ;
- regression functions (see Figure 2.1) :

(i) $h(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$

(ii) $h(x) = (\sin(2\pi x^3))^3$

- standard deviation of the error : $\sigma = \alpha R_y$; $\alpha = 0.1, 0.25, 0.5$, where R_y is the range of $h(x)$ over $x \in [0,1]$ (thus α^{-1} can be interpreted as the signal to noise ratio) ;

- regression estimators :

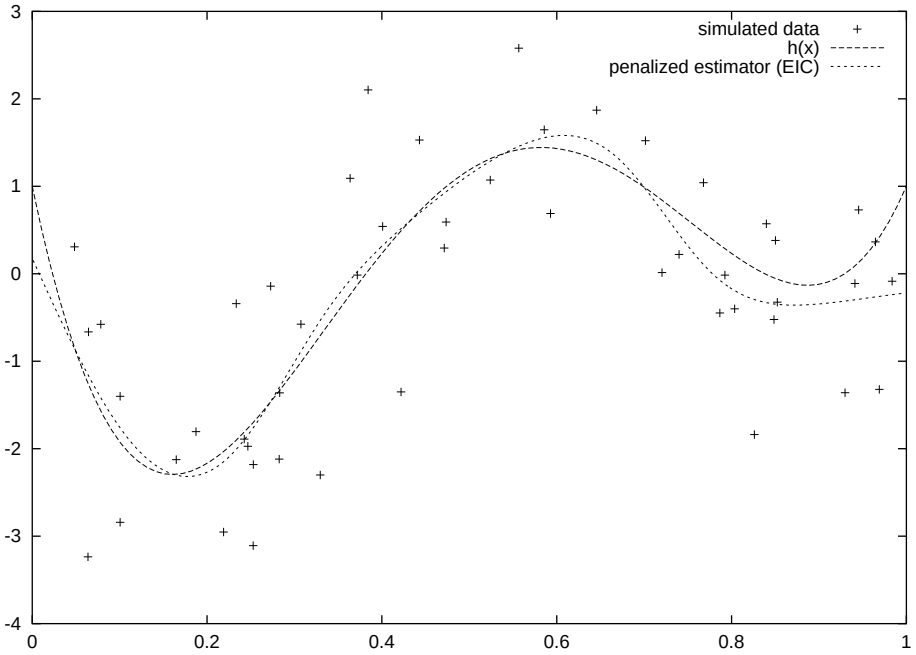
(i) polynomial estimators with $k \in \{1, 2, \ldots, 10\}$

(ii) penalized likelihood estimators, $k \in [0, +\infty[$.

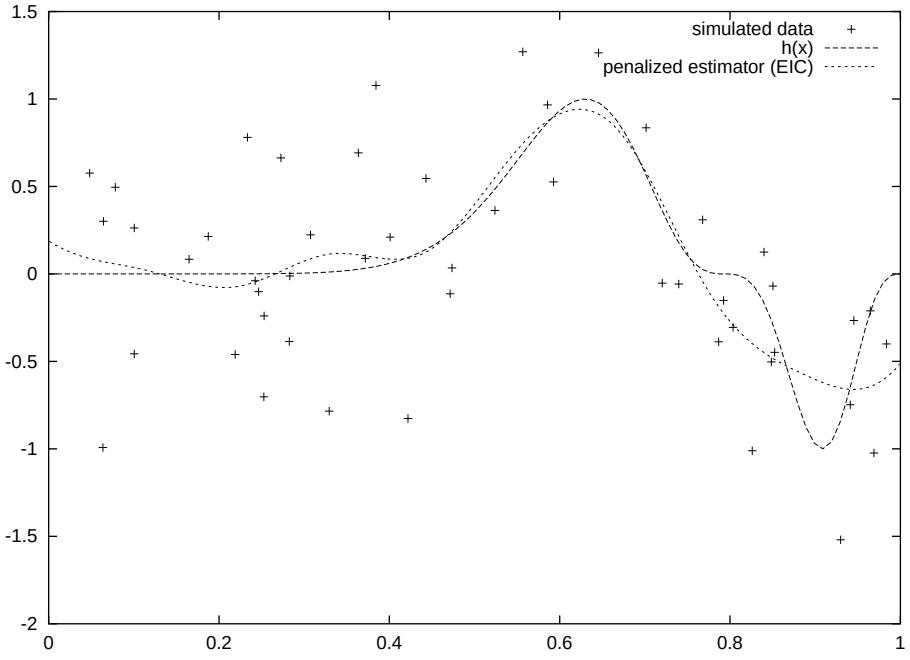
Although the exact solution of maximising the penalized likelihood is a cubic spline with knots at each observation point, this was approximated by using only 10 equidistant knots; this decreases the computation time and the approximation error is negligible in regard to the statistical error in our simulation. Figure 2.1 also displays the penalized likelihood estimate chosen by EIC.

We used 500 replications for each simulation and B=200 bootstrap replications whenever bootstrap was applied. We evaluated the performance of the selection criteria by using the true Kullback-Leibler information (or rather the informative term $\mathrm{EKL_k}(\mathcal{W})$) calculated from the true model by Monte Carlo method. High values of $\mathrm{EKL_k}(\mathcal{W})$ correspond to low Kullback-Leibler information and thus to an estimator which is close to the true density $f_{Y|X}(\cdot|\cdot)$. For each replication the smoothing parameter (k) for the penalized likelihood estimator and the degree $(k-1)$ of the polynomial estimator was chosen using each criterion, and $\mathrm{EKL_k}(\mathcal{W})$ was evaluated for each chosen estimator. Table 2.1 and 2.2 list the average $\mathrm{EKL_k}(\mathcal{W})$ for each criterion and for parametric and semi-parametric estimators. These averages must be compared to the optimal average which is the average of $\mathrm{EKL_k}(\mathcal{W})$ for estimators chosen using the true Kullback-Leibler information. For comparing the six criteria, we may note that AIC, GCV and BIC yield in many cases much lower (worse) values of $\mathrm{EKL_k}(\mathcal{W})$ than the other criteria. For the penalized likelihood estimators (Table 2.1) the differences were small between EIC, LCV and AICc, although EIC had always the highest value. For the polynomial estimators EIC achieved the highest value and the differences were more practically significant, especially for small sample sizes. AICc

FIG. 2.1 – Regression functions, simulated data, and penalized likelihood estimator chosen by EIC for two simulated cases : (a) $h(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$, (b) $h(x) = \left\{\sin\left(2\pi x^3\right)\right\}^3$ with $\alpha = 0.25$ and n=50.



(a)



(b)

tended to perform better than LCV for the polynomial model while it could give very bad values for the non-polynomial one ; indeed this could be expected since for deriving the AICc it is assumed that the true model is in the family of models considered. It may be questioned whether these differences observed between the mean $\mathrm{EKL_k}(\mathcal{W})$ computed using 500 replications really reflect differences between expectations. The standard errors of these means can be computed ; however they vary for all cases simulated and for all criteria. For instance the standard deviations of the distribution for the six first simulated cases with polynomial estimators on table 2.2 and for EIC, are respectively 0.55, 0.28, 0.17 (polynomial model) and 0.69, 0.32, 0.22 (non-polynomial model) ; the corresponding standard errors of the mean are 0.025, 0.013, 0.008, and 0.031, 0.014, 0.010. So we may consider that the observed differences between the means for EIC and LCV for these six cases (when n=30) is both statistically and practically significant since these differences are respectively : 0.20, 0.35, 0.71, 0.34, 0.18, 0.62.

Note that for polynomial estimators and for n=30, very low values of $\mathrm{EKL_k}(\mathcal{W})$ are attained by AIC, GCV, BIC, and for the non-polynomial model also by AICc : this may be explained by the fact that these criteria may select polynoms of high degree producing large extrapolation error. Such an extrapolation is needed to estimate the regression function between 0 and x_{min} and x_{max} and 1 ; for small samples the mean range of the extrapolation interval is larger than for large sample which explains why these problem occur especially for small sample. We can notice that the $\mathrm{EKL_k}(\mathcal{W})$ of the EIC is higher with polynomial estimator than with the penalized likelihood estimator when the true model is a polynomial function. On the contrary when the true model is not polynomial, the estimation method which performs the best according to the EIC is generaly the penalized likelihood estimator. Surprisingly, for $n = 100$ and high signal to noise ratio ($\sigma/R_y = 0.1$) the polynomial estimators seem to do better.

TAB. 2.1 – Average Kullback-Leibler information ($\mathrm{EKL_k}(\mathcal{W})$) for the penalized likelihood estimators for each criterion according to n, $h(\cdot)$, σ/R_y

σ/R_y	$\mathrm{EKL_k}(\mathcal{W})$ for penalized likelihood estimators						
	optimal	EIC	LCV	AIC	AICc	GCV	BIC
$h(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$ $n = 30$							
0.1	-0.93	-1.11	-1.13	-1.36	-1.11	-1.17	-1.14
0.25	-1.62	-1.72	-1.73	-2.20	-1.72	-1.79	-1.75
0.5	-2.24	-2.28	-2.29	-2.54	-2.30	-2.34	-2.30
$h(x) = \left\{\sin\left(2\pi x^3\right)\right\}^3$ $n = 30$							
0.1	-0.27	-0.51	-0.53	-4.62	-0.55	-2.83	-0.74
0.25	-0.92	-1.01	-1.02	-1.74	-1.04	-1.19	-1.07
0.5	-1.53	-1.57	-1.57	-1.71	-1.58	-1.63	-1.58
$h(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$ $n = 50$							
0.1	-0.71	-0.74	-0.77	-0.80	-0.74	-0.76	-0.75
0.25	-1.53	-1.57	-1.58	-1.60	-1.57	-1.58	-1.57
0.5	-2.19	-2.22	-2.22	-2.24	-2.22	-2.23	-2.23
$h(x) = \left\{\sin\left(2\pi x^3\right)\right\}^3$ $n = 50$							
0.1	-0.11	-0.19	-0.22	-0.54	-0.32	-0.40	-0.29
0.25	-0.86	-0.91	-0.91	-0.94	-0.91	-0.93	-0.91
0.5	-1.50	-1.52	-1.53	-1.55	-1.53	-1.54	-1.53
$h(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$ $n = 100$							
0.1	-0.58	-0.59	-0.59	-0.59	-0.59	-0.59	-0.59
0.25	-1.48	-1.49	-1.49	-1.49	-1.49	-1.49	-1.49
0.5	-2.16	-2.17	-2.17	-2.17	-2.17	-2.17	-2.18
$h(x) = \left\{\sin\left(2\pi x^3\right)\right\}^3$ $n = 100$							
0.1	-0.001	-0.032	-0.033	-0.079	-0.057	-0.072	-0.033
0.25	-0.81	-0.82	-0.83	-0.83	-0.83	-0.83	-0.82
0.5	-1.47	-1.48	-1.43	-1.48	-1.48	-1.48	-1.49

TAB. 2.2 – Average Kullback-Leibler information ($\mathrm{EKL_k}(\mathcal{W})$) for the polynomial estimator for each criterion according to n, $h(\cdot)$, σ/R_y

σ/R_y	$\mathrm{EKL_k}(\mathcal{W})$ for polynomial estimators						
	optimal	EIC	LCV	AIC	AICc	GCV	BIC
$h(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$ $n = 30$							
0.1	-0.70	-0.86	-1.06	-23.59	-0.92	-1.61	-1.06
0.25	-1.58	-1.80	-2.15	-203.7	-1.90	-190.6	-175.5
0.5	-2.23	-2.33	-3.04	-39.90	-2.50	-39.01	-8.32
$h(x) = \left\{\sin\left(2\pi x^3\right)\right\}^3$ $n = 30$							
0.1	-0.25	-0.82	-1.16	-239.9	-104.5	-173.7	-157.9
0.25	-0.93	-1.13	-1.31	-227.9	-168.8	-202.1	-178.1
0.5	-1.53	-1.60	-2.22	-40.51	-1.87	-11.73	-3.13
$h(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$ $n = 50$							
0.1	-0.59	-0.61	-0.66	-1.00	-0.65	-0.67	-0.62
0.25	-1.53	-1.57	-1.58	-1.59	-1.57	-1.58	-1.59
0.5	-2.19	-2.24	-2.25	-2.44	-2.25	-2.42	-2.26
$h(x) = \left\{\sin\left(2\pi x^3\right)\right\}^3$ $n = 50$							
0.1	-0.083	-0.28	-1.05	-2.72	-1.37	-1.66	-1.13
0.25	-0.86	-0.98	-1.08	-1.39	-1.24	-1.32	-1.07
0.5	-1.50	-1.56	-1.58	-1.79	-1.59	-1.62	-1.56
$h(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$ $n = 100$							
0.1	-0.54	-0.54	-0.55	-0.56	-0.55	-0.55	-0.54
0.25	-1.46	-1.47	-1.47	-1.48	-1.48	-1.48	-1.46
0.5	-2.15	-2.16	-2.17	-2.17	-2.16	-2.17	-2.17
$h(x) = \left\{\sin\left(2\pi x^3\right)\right\}^3$ $n = 100$							
0.1	-0.0001	-0.026	-0.049	-0.069	-0.067	-0.069	-0.046
0.25	-0.80	-0.82	-0.84	-0.85	-0.85	-0.85	-0.84
0.5	-1.47	-1.50	-1.50	-1.50	-1.50	-1.50	-1.51

2.5 Illustration

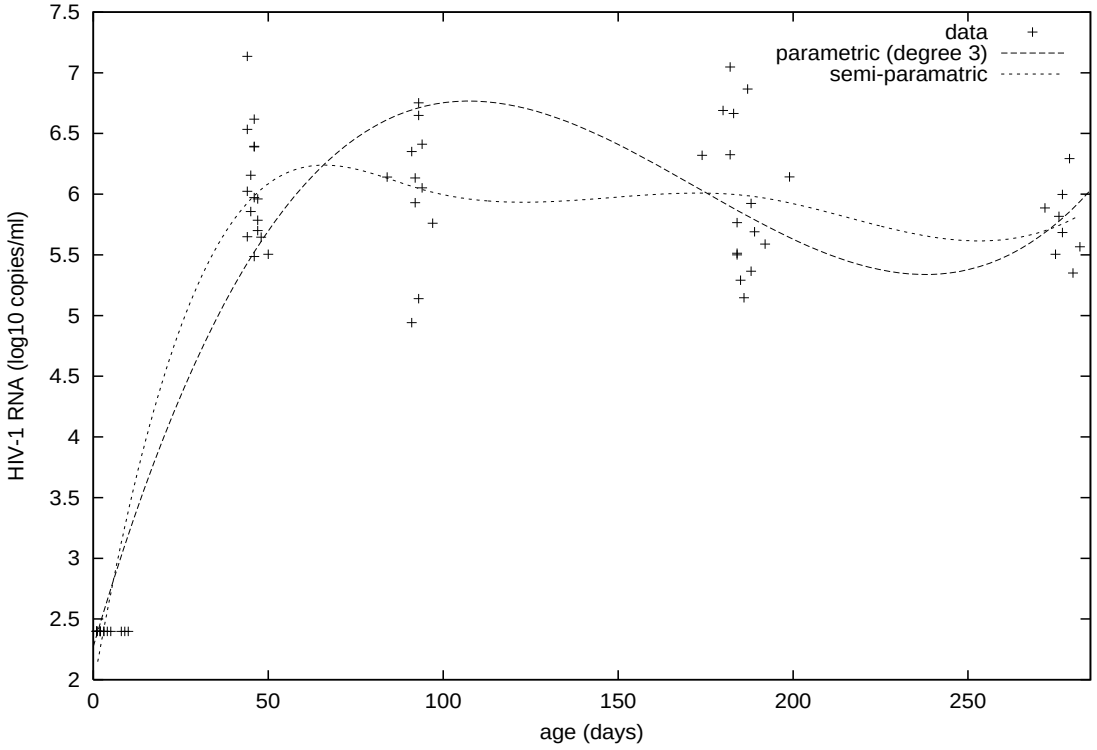
This section illustrates the use of the EIC criterion to select a model estimating the dynamics of HIV-1 RNA levels in infected children as a function of age ; the data come from The Ditrame ANRS 049 trial. This trial was designed to assess the efficacy of a maternal short ZDV regimen to reduce mother-to-child transmission in Abidjan, Côte d'Ivoire (Dabis et al., 1999). The children were not treated, and were followed longitudinally from birth. It was planned to obtain plasma samples during the first week of life and at day 45, 90, 180, 270 ; however many children had some missing values. We focuss on a group of infants infected early in the peripartum period, but not in utero. This group is defined as having a PCR (Polymerase Chain Reaction) during the first week and a positive one at 45 days. Among this group we selected a subgroup of 17 infants still alive at 270 days and presenting 4 measurements. We analysed the logarithm of the RNA levels, which have a distribution closer to the normal ; for the first week where HIV-1 RNA was undetectable, we attributed the threshold value (2.38 log10 copies/ml). We represent the longitudinal data as multivariate data : for each subject we observe (Y_i, X_i) where $Y_i = (Y_{i1}, Y_{i2}, Y_{i3}, Y_{i4})'$ and $X_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})'$ and the conditional distribution of Y_i given X_i was specified by :

$$Y_{ij} = h(X_{ij}) + \epsilon_{ij}$$

with $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ and an AR1 correlation structure (that is $\mathrm{cor}(\epsilon_{ij}, \epsilon_{ik}) = e^{-\gamma|X_{ij} - X_{ik}|}$). If the design were fixed with the same X_i for each infant, the (Y_i, X_i) would be i.i.d (albeit with a degenerate distribution). Here the X_i deviate from the fixed design by missing data and approximate days of visits according to a pattern that we consider independent from one child to another. We applied the EIC to select the smoothing parameter for the penalized likelihood estimator, and the degree of the polynomial estimator. Figure 2.2 represents the parametric and the semi-parametric models chosen.

According to the EIC, the best polynomial fitting is of degree 3, with EIC=-63.06. The penalized likelihood estimator selected obtains EIC=-54.15. Therefore, we conclude that the best representation of the viral load dynamic is by using penalized likelihood

FIG. 2.2 – Evolution of HIV-1 RNA in children infected peripartum or early postanal, parametric and semi-parametric models



estimator, which intuitively also seems the most satisfactory when looking at Figure 2.2. The mean HIV-1 RNA levels seems to increase rapidly until a peak about 2 months after birth, and then slowly decrease from 6.2 log10 copies/ml at two month down to 5.6 log10 copies/ml at nine. The rapid increase is not surprising according to the children selected : they all have a HIV-1 RNA level under the threshold at birth and a positive one at 45 days.

2.6 Discussion

Our simulation results show that EIC is better than its competitors at least for small sample size. For large sample size the criteria tend to be equivalent and indeed several results of asymptotic equivalence have been proved (Shibata, 1997). However for large sample size we might also consider more complicated models leading to richer families of

estimators and hence more difficult model choice problems. For instance we may consider a multivariate problem in which Y_{ij} would have a distribution depending on $h_m(X_{im}), m = 1, \ldots, p$ and the problem would be the choice between several estimators of $h_m(X_{im})$. For instance Y_{ij} could be repeated binary observations and the model could be a logistic model with additive structure for the effect of covariates X_m, in the spirit of Hastie and Tibshirani (1990). In such complicated problems EIC has a potential of being better than competitors. It must be noted that the theory we have developed allows treatment of this kind of problem while AIC has been applied only to parametric models or to simple semi-parametric ones such as the regression problem treated in our simulation. Among the competitors of EIC, only LCV can be used in full generality. Further work is needed to develop EIC to allow treatment of problems with incomplete data.

From a computationnal point of view, EIC requires more (resp. less) computations than LCV when the number of replicates is larger (resp. smaller) than the sample size. In practice for the search of the index k for EIC we keep the same B bootstrap samples for each value of k : this spares some computation time and avoids some variability when comparing values of EIC for different values of k. Computation times remain rather short ; for instance it took 11.3 seconds of CPU time on a pentium IV to obtain the penalized likelihood estimator chosen by EIC shown in figure 2.1 (b).

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* **52**, 345–370.

Dabis, F., Msellati, P., Meda, N., Wellfens-Ekra, C., You, B., Manigart, O., Leroy, V., Simonon, A., Cartoux, M., Combe, P., Ouangre, A., Ramon, R., Ky-Zerbo, O., Montcho, C., Salamon, R., Rouzioux, C., Van de Perre, P., L., M. and Group, D. S. (1999). 6-month efficacy, tolerance, and acceptability of a short regimen of oral zidovudine to reduce vertical transmission of hiv in breastfed children in Côte d'Ivoire and Burkina Faso: a double-blind placebo-controlled multicentre trial. *Lancet* **353**, 786–792.

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman and Hall, London.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hurvich, C. M., Simonoff, J. and Tsai, C. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society B* **60**, 271–293.

Hurvich, C. M. and Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Ann. Inst. Statist. Math* **49**, 411–434.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika* **83**, 875–890.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Statist. Comput* **9**, 363–379.

Pan, W. (1999). Bootstrapping likelihood for model selection with small samples. *Journal of Computational and Graphical Statistics* **8**, 687–698.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Shibata, R. (1997). Bootstrap estimate of Kullback-Leibler information for model selection. *Statistica Sinica* **7**, 375–394.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society B* **46**, 1–52.

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods* **A7**, 13–26.

Chapitre 3

Modélisation du risque de mésothéliome pleural lié à une exposition professionnelle à l'amiante

3.1 Contexte

Le mésothéliome est une tumeur maligne le plus souvent localisée au niveau de la plèvre et dont l'amiante est pratiquement la seule étiologie établie à ce jour. Cette tumeur, à caractère essentiellement professionnel, concerne principalement les hommes. Elle survient généralement après un délai de 30 à 40 ans après le début de l'exposition. Du fait de l'intensification de l'utilisation de l'amiante depuis le début du siècle dans différents secteurs industriels, l'incidence de cette pathologie jusqu'alors très rare est en forte augmentation depuis les années 1950-1960 dans les pays industrialisés [8].

Le Programme National de Surveillance du Mésothéliome (PNSM) a été initié en 1998. Il s'agit d'une action concertée faisant appel à des spécialistes de divers domaines : épidémiologie, pneumologie, cancérologie, anatomopathologie, médecine du travail et hygiène industrielle. Le PNSM est organisé en cinq volets : Incidence, Étiologie, Anatomopatholo-

gie, Clinique et Médico-social, dont la coordination est assurée par le Département Santé Travail (DST) de l'Institut de Veille Sanitaire (InVS).

Outre l'estimation de l'incidence nationale du mésothéliome pleural en France et l'étude de la reconnaissance de la maladie comme maladie professionnelle, l'un des objectifs principaux du PNSM est d'estimer la relation dose-effet entre l'exposition à l'amiante et le risque de survenue de la maladie. Une enquête étiologique cas-témoins a été mise en place par le Laboratoire Santé Travail et Environnement (LSTE) à partir de 1998 dans 17 départements (étendue aujourd'hui à 19 départements).

Pour chaque cas, une sélection de deux témoins est effectuée en population générale (sujets indemnes de la maladie), appariés sur le département de domicile, l'âge et le sexe. Un questionnaire ad-hoc est administré auprès de chaque sujet vivant. L'expertise du calendrier professionnel par le LSTE permet d'évaluer l'exposition professionnelle à l'amiante et de calculer pour chacun des sujets une dose cumulée d'exposition exprimée en fibres/millilitre-années (f/ml-années).

Afin d'étudier la relation dose-effet, plusieurs méthodes paramétriques et non-paramétriques ont été proposées et réalisées, entraînant un problème de choix de modèles non résolu avec les critères statistiques usuels.

3.2 Objectif

Ma contribution à ce travail a été de proposer et d'appliquer un critère d'information permettant de faire un choix entre plusieurs modélisations. Apres avoir décrit les données, nous définissons le critère d'information EIC [11] dans ce contexte. Les différentes modélisations effectuées sont décrites dans la section 3.5. Les résultats sont présentés dans la section 3.6. Enfin nous présentons en section 3.7 une autre approche de sélection de modèles proposée par Birgé et Massart [2] et nous l'appliquons à ces données.

3.3 Les données

En raison du caractère de la maladie, nous nous sommes intéressés au risque de mésothéliome chez les hommes en milieu professionnel. Pour l'analyse, l'échantillon comprenait 374 sujets (174 cas et 200 témoins). La variable d'intérêt est la maladie (variable dichotomique : malade-non malade). La variable explicative étudiée est l'exposition professionnelle à l'amiante (variable continue : dose cumulée d'exposition mesurée en f/ml-années).

Nous observons donc un échantillon $\mathcal{W} = (w_1, \ldots, w_n)$ avec $w_i = (y_i, x_i), (y_i \in \mathbb{N}, x_i \in \mathbb{R})$. Chaque w_i est une réalisation i.i.d. d'une variable aléatoire $W_i = (Y_i, X_i), (Y_i \in \mathbb{N}, X_i \in \mathbb{R})$ de distribution $F_W(.,.) = F_{Y,X}(.,.)$. Y_i représente la présence de la maladie ($Y_i = y^0 = 0$ non malade ; $Y_i = y^1 = 1$ malade) et X_i définit le niveau d'exposition. Notre but étant d'estimer la relation dose-effet, nous nous sommes intéressés à la distribution conditionnelle de Y sachant $X = x$ définie par la distribution $P_{Y|X}(.|X = x)$, qui est une distribution de bernouilli d'espérance $p(x)$. Nous considérons $\widehat{Q}^K_{Y|X}(\cdot|\cdot; \mathcal{W})$ un estimateur de $P_{Y|X}(\cdot|\cdot)$ construit à partir de $\mathcal{W}$. $\widehat{Q}^K_{Y|X}(\cdot|\cdot; \mathcal{W})$ est une distribution de Bernouilli d'espérance $\widehat{q}^K(x; \mathcal{W})$. Le paramètre K représente soit un paramètre de lissage dans une approche non-paramétrique, soit la dimension de l'espace dans une approche paramétrique.

Remarque : le fait que les données proviennent d'une étude cas-témoins pose un problème théorique car les observations sont sélectionnées. La distribution de W_i dont il est question est en fait la distribution conditionnelle à la sélection.

3.4 Critère d'Information

Nous nous proposons d'utiliser le critère EIC dans le choix d'estimateurs de distribution de lois discrètes. Le critère EIC a été présenté dans un cadre de sélection d'estimateur de densité de variable continue. Dans le cas présent, notre variable d'intérêt est discrète, la dérivation du EIC est légèrement différente mais le critère reste identique.

Nous définissons l'information de Kullback-Leibler mesurant la perte d'information lorsque $\widehat{Q}^K_{Y|X}(\cdot|\cdot;\mathcal{W})$ est utilisée pour approcher la vraie distribution $P_{Y|X}(\cdot|\cdot)$. Pour un X donné nous avons :

$$\begin{aligned} \mathrm{I}\left\{P_{Y|X}(.|X);\widehat{Q}^K_{Y|X}(.|X;\mathcal{W})\right\} &= \sum_{i=0}^{1} P_{Y|X}(y^i|X)\log P_{Y|X}(y^i|X) \\ &\quad - \sum_{i=0}^{1} P_{Y|X}(y^i|X)\log \widehat{Q}^K_{Y|X}(y^i|X;\mathcal{W}). \end{aligned}$$

Dans une procedure de sélection, seul le second terme est pertinent. On note donc

$$\mathrm{KL_K}(X,\mathcal{W}) = \mathrm{E}\left\{\log \widehat{Q}^K_{Y|X}(Y|X;\mathcal{W})|X,\mathcal{W}\right\}$$

où $(Y,X) = W$ est une nouvelle observation et $W \sim F_{Y,X}$, indépendemment de l'échantillon $\mathcal{W}$. Comme on veut que notre critère ne dépende pas de X, nous définissons

$$\mathrm{EKL_K}(\mathcal{W}) = \mathrm{E}\left\{\mathrm{KL_K}(X,\mathcal{W})|\mathcal{W}\right\} = \mathrm{E}\left\{\log \widehat{Q}^K_{Y|X}(Y|X;\mathcal{W})|\mathcal{W}\right\}.$$

Un estimateur naturel est trouvé en remplacant $F_{Y,X}$ par sa distribution empirique $\widehat{F}_{Y,X}$:

$$\begin{aligned} \widehat{\mathrm{EKL}}_K(\mathcal{W}) &= \frac{1}{n}\sum_{i=1}^{n} \log \widehat{Q}^K_{Y|X}(Y_i|X_i;\mathcal{W}) \\ &= \frac{1}{n}\sum_{i=1}^{n} \left[Y_i \log \widehat{q}^K(X_i;\mathcal{W}) + (1-Y_i)\log\left\{1-\widehat{q}^K(X_i;\mathcal{W})\right\}\right] \end{aligned}$$

On peut reconnaitre (1/n) fois la log-vraisemblance de $\mathcal{W}$ pour la probabilité conditionnelle specifiée par la distribution $\widehat{Q}^K_{Y|X}(\cdot|\cdot;\mathcal{W})$.
Cet estimateur est généralement biaisé. Le calcul du biais par la technique du bootstrap

est expliquée par Liquet et Commenges [11]. Le résultat est explicité dans ce cas :

$$\begin{aligned}\widehat{b}(\mathcal{W}) &\simeq \frac{1}{B}\sum_{j=1}^{B}\left\{\frac{1}{n}\sum_{i=1}^{n}\log \widehat{Q}^{K}_{Y|X}(Y_i^j|X_i^j;\mathcal{W}^j) - \frac{1}{n}\sum_{i=1}^{n}\log \widehat{Q}^{K}_{Y|X}(Y_i|X_i;\mathcal{W}^j)\right\} \\ &\simeq \frac{1}{B}\sum_{j=1}^{B}\left\{\frac{1}{n}\left[\sum_{i=1}^{n}\left[Y_i^j \log \widehat{q}^{K}(X_i^j;\mathcal{W}^j) + (1-Y_i^j)\log\left(1-\widehat{q}^{K}(X_i^j;\mathcal{W}^j)\right)\right.\right.\right. \\ &\qquad \left.\left.\left. -Y_i \log \widehat{q}^{K}(X_i;\mathcal{W}^j) - (1-Y_i)\log\left\{1-\widehat{q}^{K}(X_i;\mathcal{W}^j)\right\}\right]\right]\right\}\end{aligned}$$

où $\mathcal{W}^j$ représente les répliques bootstrap ($\mathcal{W}^j \sim \widehat{F}_{Y,X}$) et B est le nombre de répliques. En définissant $\mathcal{L}^{\widehat{Q}^{\mathcal{W}^j}}(\mathcal{W})$ la vraisemblance de l'échantillon $\mathcal{W}$ pour la probabilité conditionnelle spécifiée par $\widehat{Q}^{K}_{Y|X}(\cdot|\cdot;\mathcal{W}^j)$, le biais s'écrit plus simplement :

$$\widehat{b}(\mathcal{W}) \simeq \frac{1}{B}\frac{1}{n}\sum_{j=1}^{B}\left\{\log \mathcal{L}^{\widehat{Q}^{\mathcal{W}^j}}(\mathcal{W}^j) - \log \mathcal{L}^{\widehat{Q}^{\mathcal{W}^j}}(\mathcal{W})\right\}$$

Finalement, le critère est :

$$\text{EIC} = \frac{1}{n}\log \mathcal{L}^{\widehat{Q}^{\mathcal{W}}}(\mathcal{W}) - \widehat{b}(\mathcal{W}).$$

Nous choisirons l'estimateur $\widehat{Q}^{K}_{Y|X}(\cdot|\cdot;\mathcal{W})$ qui maximise le EIC.

3.5 Les estimateurs

Nous présentons les différents estimateurs proposés pour estimer la relation dose-effet. Nous les classons en deux familles : paramétriques pour les constantes par morceaux et les polynômes fractionnels et non-paramétriques pour la vraisemblance pénalisée et la regression locale pondére (LOESS). Pour tous ces estimateurs, la distribution de $Y|X = x$ est une distribution de Bernouilli d'espérance inconnue. C'est cette espérance qui est modélisée. Le modèle classique en épidémiologie est le modèle logistique [6] :

$$\text{Logit}\{pr(Y_i = 1|X_i = x)\} = \text{Logit } p(x) = f(x)$$

ou encore

$$pr(Y_i = 1|X_i = x) = p(x) = \frac{e^{f(x)}}{1 + e^{f(x)}}$$

Les différentes distributions seront définies par les modélisations de $p(\cdot) = pr(Y_i|X_i = x)$ ou encore de $f(\cdot)$.

3.5.1 Famille paramétrique

Notation : on note $\tilde{\mathcal{W}} = (w_{(1)}, \ldots, w_{(n)})$ puis $\{w_{(i)} = (y_{(i)}, x_{(i)})\}$ l'échantillon ordonné par la valeur de l'exposition. Cet échantillon se décompose en deux échantillons $\tilde{\mathcal{W}}^0_{n_0}$ et $\tilde{\mathcal{W}}^1_{n_1}$. $\tilde{\mathcal{W}}^0_{n_0}$ représente l'échantillon pour les non-exposés ($X_i = 0$) et $\tilde{\mathcal{W}}^1_{n_1}$ l'échantillon des individus exposés à l'amiante. Au moment de l'analyse, l'analyse est composée de $n_0 = 117$ et $n_1 = 257$.

$$\tilde{\mathcal{W}} = \Big(\underbrace{(y_{(1)}, x_{(1)}), \ldots, (y_{(n_0)}, x_{(n_0)})}_{\tilde{\mathcal{W}}^0_{n_0} = (\tilde{Y}^0_{n_0}, \tilde{X}^0_{n_0})}, \underbrace{(y_{(n_0+1)}, x_{(n_0+1)}), \ldots, (y_{(n)}, x_{(n)})}_{\tilde{\mathcal{W}}^1_{n_1} = (\tilde{Y}^1_{n_1}, \tilde{X}^1_{n_1})} \Big)$$

Constante par morceaux

La variable d'exposition est considérée comme une variable catégorielle. Au vue de la distribution de la variable d'exposition, nous avons choisi de considérer les modèles

suivants :

$$f^K(x_i) = \beta_1 \mathbb{1}_{x_i \in I_1} + \beta_2 \mathbb{1}_{x_i \in I_2} + \ldots + \beta_K \mathbb{1}_{x_i \in I_K}$$

où les I_k sont définis de la façon suivante :

$$m = \bigcup_{k=1}^{K} I_k$$

et forment une partition m de l'étendue de la variable d'exposition telle que

$$\begin{cases} I_k =]\tau_{k-1}, \tau_k] \text{ pour } \; k = 2, \ldots, K \;\; \tau_0 = 0, \;\; \tau_K = x_{(n)} \text{ et } I_1 = [\tau_0, \tau_1] \text{ et } \tau_1 = 0 \\ I_k \bigcap I_{k'} = \emptyset \;\; \text{pour} \;\; k = 1, \ldots, K \;\; \text{et} \;\; k' = 1, \ldots, K \end{cases}$$

où τ_k avec $k = 2, \ldots, K-1$ représente les points de coupure de la variable d'exposition $\tilde{X}^1_{n_1}$ (individus exposés). On note $D_m = K$ la dimension de la partition m. La stratégie la plus usuelle en épidémiologie est de choisir la médiane pour un point de coupure ($D_m = 3$) et de prendre le premier et le second tercile comme points de coupure pour une partition de dimension 3 ($D_m = 4$), et ainsi de suite. Cette stratégie est résumée dans le tableau 3.1 ci-dessous.

TAB. 3.1 – Stratégie du choix des points de coupures

K=D_m	τ_2	τ_3	τ_4	τ_5	...	τ_8
3	médiane					
4	1^{er}tercile	2^{nd}tercile				
5	1^{er}quartile	$2^{ième}$quartile	$3^{ième}$quartile			
6	1^{er}quintile	$2^{ième}$quintile	$3^{ième}$quintile	$4^{ième}$quintile		
⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	$1^{ième}$octile	$2^{ième}$octile	$3^{ième}$octile	$4^{ième}$octile	...	$7^{ième}$octile

Cette modélisation implique donc une probabilité différente entre les individus faiblement exposés et les non-exposés. L'hypothèse implique l'existence d'un risque de mésothéliome pour des individus faiblement exposés. Cette modélisation permet de comparer des indi-

vidus exposés à une certaine dose d'amiante aux individus non-exposés. D'autres partitions de m auraient pu être retenues mais le choix des différentes classes (ou points de coupure) devient alors problématique. Finalement pour cette modélisation, l'estimateur par morceaux retenu parmi $\left\{\widehat{f}^K(\cdot) \quad K = 3, \ldots, 9\right\}$ sera celui qui maximise le EIC. $\left\{\widehat{f}^K(\cdot) \quad K = 3, \ldots, 9\right\}$ correspond aux estimateurs des modèles $\left\{f^K(\cdot) \quad K = 3, \ldots, 9\right\}$ où les différents paramètres sont estimés par le principe du maximum de vraisemblance.

Polynôme fractionnel

Les polynômes fractionnels [12] sont des polynômes où les puissances en x peuvent être non entières ou négatives. Cette modélisation permet d'avoir plus d'information qu'en utilisant des constantes par morceaux mais elle est plus difficile à interpréter. Nous utiliserons la même démarche que Lemeshow [6]. Le degré d'un polynôme fractionnel est défini comme le nombre de termes en puissance de x, noté d. Par exemple :

$$\begin{aligned} f^k(x) &= \beta_0 + \beta_1 x^{-1} & d = 1 \\ f^k(x) &= \beta_0 + \beta_1 x^{-1} + \beta_2 x^2 & d = 2 \end{aligned}$$

On note les puissances d'un polynôme fractionnel $p_1, p_2, \ldots$. Le vecteur des puissances est noté par $\mathbf{p}$. Dans le premier exemple $\mathbf{p} = -1$ et dans le second exemple $\mathbf{p} = (-1, 2)$. Nous choisissons de nous limiter aux polynômes de degré 1 et 2 puisqu'ils fournissent déja une grande variété de modèles. Les polynômes seront choisis en incluant les puissances p_i appartenant à l'ensemble :

$$\mathcal{P} = \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$$

La puissance de 0 correspond au $\ln x$. Lorsque $d > 1$ et p_1=p_2, le polynôme fractionnel est

$$f^k(x) = \beta_0 + \beta_1 x^{p_1} + \beta_2 x^{p_1} \ln x$$

Nous donnons quelques exemples de polynômes fractionnels dans le tableau ci dessous :

TAB. 3.2 – Exemples de polynômes fractionnels

Puissances	Modèle
(0,0)	$\beta_0 + \beta_1 \ln x + \beta_2 (\ln x)^2$
(0.5,0.5)	$\beta_0 + \beta_1 \sqrt{x} + \beta_2 \sqrt{x} \ln x$
(2,2)	$\beta_0 + \beta_1 x^2 + \beta_2 x^2 \ln x$
(3,3)	$\beta_0 + \beta_1 x^3 + \beta_2 x^3 \ln x$
(0,3)	$\beta_0 + \beta_1 \ln x + \beta_2 x^3$
(-2,-2)	$\beta_0 + \beta_1 (1/x^2) + \beta_2 (1/x^2) \ln x$
(1,1)	$\beta_0 + \beta_1 x + \beta_2 x \ln x$

Le meilleur modèle de degré d=1 est celui qui a le plus grand EIC parmi les huit polynômes essayés :

$$f^k(x) = \beta_0 + \beta_1 x^{p_1}$$

avec $p_1 \in \{-2, -1, -0.5, 0, 0.5, 1, 2, 3\}$

Le meilleur modèle de degré d=2 est celui qui a le plus grand EIC parmi les 64 polynômes essayés en choisisant toutes les combinaisons possibles de p_1 et p_2 appartenant à $\mathcal{P}$:

$$f^k(x) = \beta_0 + \beta_1 x^{p1} + \beta_2 x^{p2}.$$

Finalement le polynôme fractionnel retenu sera celui qui a l'EIC le plus élevé entre le meilleur de degré 1 et le meilleur de degré 2.

3.5.2 Famille non-paramétrique

Estimateur par vraisemblance pénalisée

Une approche plus souple est d'imposer à la fonction $f(\cdot)$ d'être continue et d'avoir de faibles variations locales. Pour cela, un moyen est de pénaliser la vraisemblance par

un terme qui est d'autant plus grand que la fonction $f(\cdot)$ est peu lisse. La pénalisation choisie est $\int f''^2$. La fonction f doit donc, dans ce cas, appartenir à la classe des fonctions continues, deux fois différentiables et dont la dérivée seconde est de carré intégrable. Pour cet estimateur, $\widehat{q}^k = \dfrac{e^{\widehat{f}^k(x)}}{1+e^{\widehat{f}^k(x)}}$ maximise la vraisemblance pénalisée [7] :

$$\sum_{i=1}^{n} [Y_i \log p(X_i) + (1-Y_i) \log (1-p(X_i))] - k \int f''(u)du \tag{3.1}$$

où k représente le paramètre de lissage. k contrôle l'équilibre entre l'ajustement aux données et la régularité de la fonction. Il sera choisi en maximisant l'EIC. Nous proposons d'utiliser une base de fonctions splines pour obtenir une approximation de l'estimateur de l'espérance de Y qui n'est connue qu'implicitement comme le maximum de l'expression (3.1).

Régression locale pondérée

Nous proposons d'utiliser la méthode LOESS ("locally weighted scatterplot smother") développée par Cleveland [5]. Il s'agit de la régression locale pondérée . Le principe de cette méthode est de "laisser les données parler d'elles mêmes". Cette méthode permet d'obtenir un estimateur lisse $\widehat{q}^k(x)$ de $p(x)$. Dans une fenêtre mobile d'ordre k, un point lissé est calculé en le prenant sur la droite de régression linéaire (les moindres carrés sont pondérés). Cette dernière étant calculée avec les k points de la fenêtre. Le degré du polynôme de régression (noté d) peut être différent de 1 (régression linéaire). Le calcul s'effectue dans chaque fenêtre en affectant des poids différents à chacun des points. Les poids sont définis de la façon suivante : Soit $\Delta_i(x)$ la distance euclidienne de x à $X_i = x_i$. Soit $\Delta_{(i)}(x)$ les valeurs de ces distances ordonnées de façon croissante, et notons

$$T(u) = \begin{cases} (1-u^3)^3 & \text{pour } 0 \leq u < 1 \\ 0 & \text{sinon} \end{cases}$$

la fonction de poids tricubique. On définit un poids pour (y_i, x_i) par

$$w_i(x) = T\left(\frac{\Delta_i(x)}{\Delta_{(k)}(x)}\right).$$

Les w_i décroissent quand la distance de x_i à x croit. Le lissage dépend de l'ordre de la fenêtre mobile k. Ce paramètre, ou de façon équivalente la proportion $\alpha = \frac{k}{n}$, sera déterminée et choisie en maximisant le EIC. Nous modélisons directement $p(x)$ et non $f(x)$ pour des raisons numériques pour l'application de l'EIC. De plus, pour s'assurer que $\widehat{q}^k(x) \in [0:1]$, nous choisissons de se limiter à prendre un degré de polynôme $d = 0$. Finalement notre estimateur par LOESS est une moyenne mobile pondérée :

$$\widehat{q}^k(x) = \frac{\sum_{i=1}^n w_i(x) Y_i}{\sum_{i=1}^n w_i(x)}.$$

3.6 Résultats

Tous les tableaux de résultats présentent le biais $\widehat{b}(\mathcal{W})$ nécessaire au calcul du EIC et son coefficient de variation $CoVa = \frac{\sigma_{\widehat{b}(\mathcal{W})}}{\widehat{b}(\mathcal{W})}$. Ce terme représente la qualité d'approximation du calcul par bootstrap du biais.

3.6.1 Constante par morceaux

Le tableau 3.3 présente les résultats des estimateurs par morceaux pour le critère EIC et AIC.

TAB. 3.3 – Valeurs des critères pour les estimateurs par morceaux

dimensions	τ	vraisemblance	AIC	-2EIC	biais	CoVa
3	médiane	-207.82	421.65	422.35	3.35	0.10
4	terciles	-207.41	422.81	423.26	4.23	0.08
5	**quartiles**	**-203.30**	**416.60**	**417.51**	**5.46**	**0.07**
6	quintiles	-204.29	420.58	421.5	6.47	0.06
7	sextiles	-203.76	421.52	422.89	7.69	0.05
8	septiles	-200.41	416.82	417.76	8.47	0.07
9	octiles	-202.40	422.80	420.56	7.88	0.09

Ce tableau présente -2EIC pour pouvoir les comparer aux AIC. L'estimateur par morceaux ayant le EIC maximal (ou -2EIC minimum) est donc l'estimateur par morceaux de dimension 5 avec les differents quartiles comme points de coupure. Avec le AIC, nous parvenons aux même conclusions.

3.6.2 Polynôme fractionnel

Le tableau 3.4 présente uniquement le résultat des 4 meilleurs polynômes fractionnels de dimension 1 ainsi que les 4 meilleurs polynômes fractionnels de dimension 2 au sens de l'EIC. Le polynôme fractionnel $\mathbf{p} = (\mathbf{0})$ est le polynôme ayant un EIC maximal (-2EIC minimal). Le AIC pour ce polynôme est aussi minimal.

TAB. 3.4 – Valeurs des critères pour les polynômes fractionnels

puissance	modèle	vraisemblance	AIC	-2EIC	biais	CoVa
(-0.5)	$\beta_0 + \beta_1 x^{-0.5}$	-223.29	450.59	452.03	2.72	0.10
(0)	$\boldsymbol{\beta_0 + \beta_1 \ln(x)}$	**-204.27**	**412.54**	**413.39**	**2.42**	**0.14**
(1)	$\beta_0 + \beta_1 x$	-221.87	447.75	449.90	3.08	0.12
(2)	$\beta_0 + \beta_1 x^2$	-240.20	484.40	489.82	4.71	0.09
(-0.5,0)	$\beta_0 + \beta_1 \ln(x) + \beta_2 x^{-0.5}$	-203.46	412.93	413.63	3.35	0.10
(0,0)	$\beta_0 + \beta_1 \ln(x) + \beta_2 (ln(x))^2$	-207.07	413.28	414.14	3.43	0.10
(0,0.5)	$\beta_0 + \beta_1 \ln(x) + \beta_2 x^{0.5}$	-204.13	414.26	415.37	3.55	0.10
(0,1)	$\beta_0 + \beta_1 \ln(x) + \beta_2 x$	-204.21	414.42	415.48	3.53	0.10

3.6.3 Vraisemblance pénalisée

La modélisation par vraisemblance pénalisée ne donne pas des résultats satisfaisants (voir figure 3.1 et 3.2). La pénalisation par $\int f''^2$ n'est apparemment pas adaptée à ces données. Cette pénalisation ne convient pas aux fonctions à fortes variations locales. L'échantillon comprend plusieurs cas de mésothéliome exposés à de très faibles doses d'amiante ce qui engendre un saut de la probabilité pour des faibles doses.

3.6.4 Régression locale pondérée

Pour la régression locale pondérée, le tableau 3.5 liste le resultat pour quelques valeurs de α. La meilleure régression locale pondérée au sens de l'EIC est réalisée pour $\alpha = 0.35$.

TAB. 3.5 – Valeurs du EIC pour la régression locale pondérée

$\alpha = \frac{k}{n}$	vraisemblance	-2EIC	biais	CoVa
0.35	**-201.38**	**418.74**	**7.99**	**0.06**
0.40	-203.78	422.33	7.38	0.06
0.50	-204.47	421.19	6.13	0.08
0.60	-209.92	429.57	4.86	0.09
0.70	-215.41	438.64	3.90	0.10
0.80	-219.79	446.02	3.23	0.12

3.6.5 Conclusion

La meilleure modélisation, au sens de l'EIC, est obtenue par le modèle avec des polynômes fractionnels (voir le tableau 3.6 récapitulatif ci-dessous).

TAB. 3.6 – Valeurs du EIC pour les meilleures modélisations

modèle	-2xEIC
constant par morceaux	417.51
polynôme fractionnel	413.39
loess	418.74

La figure 3.1 représente la modélisation de $p(x) = pr(Y_i = 1|X_i = x)$ par les meilleurs estimateurs. Nous représentons aussi à la figure 3.2, le rapport de côtes (odds ratio) pour les différents estimateurs. L'odd ratio représente l'association entre l'exposition et la maladie. Il est défini par :

$$OR(x) = \frac{pr(Y = 1|X = x)/(1 - pr(Y = 1|X = x))}{pr(Y = 1|X = 0)/(1 - pr(Y = 1|X = 0))}$$

L'estimateur par morceaux donne un résultat proche du meilleur estimateur (EIC=417.51

pour l'estimateur par morceau et 413.39 pour le polynôme fractionel). De plus cet estimateur est souvent plus apprécié chez les épidémiologistes pour sa facilité d'interprétation. C'est pour ces raisons que nous allons étudier plus précisement cette modélisation. Nous proposons d'appliquer la méthode de sélection de modèles de Birgé-Massart qui permet d'étudier toutes les partitions possibles et de sélectionner la meilleure en se basant sur un critère pénalisé.

FIG. 3.1 – Représentation de $(p(x), x)$ pour les meilleurs estimateurs

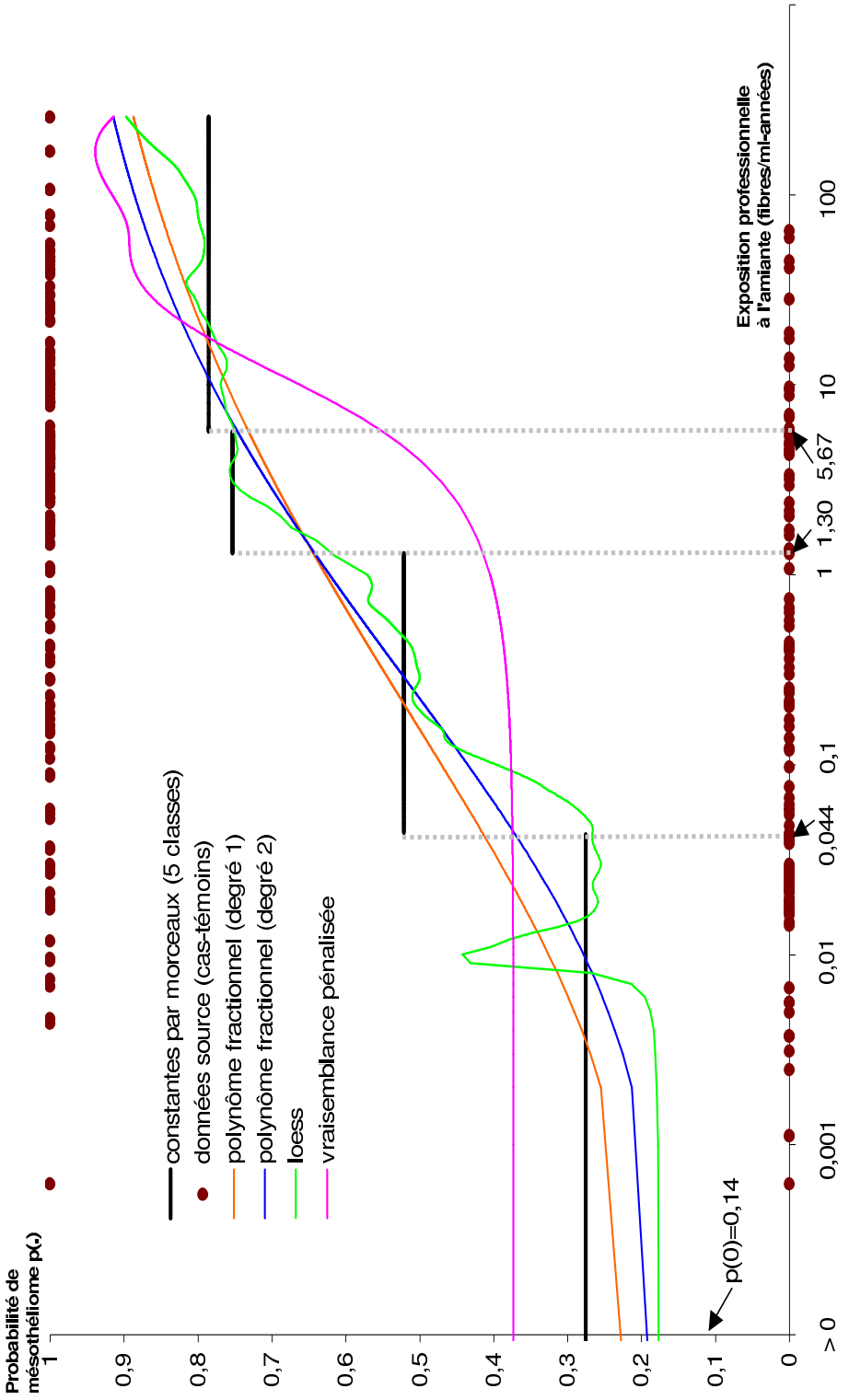
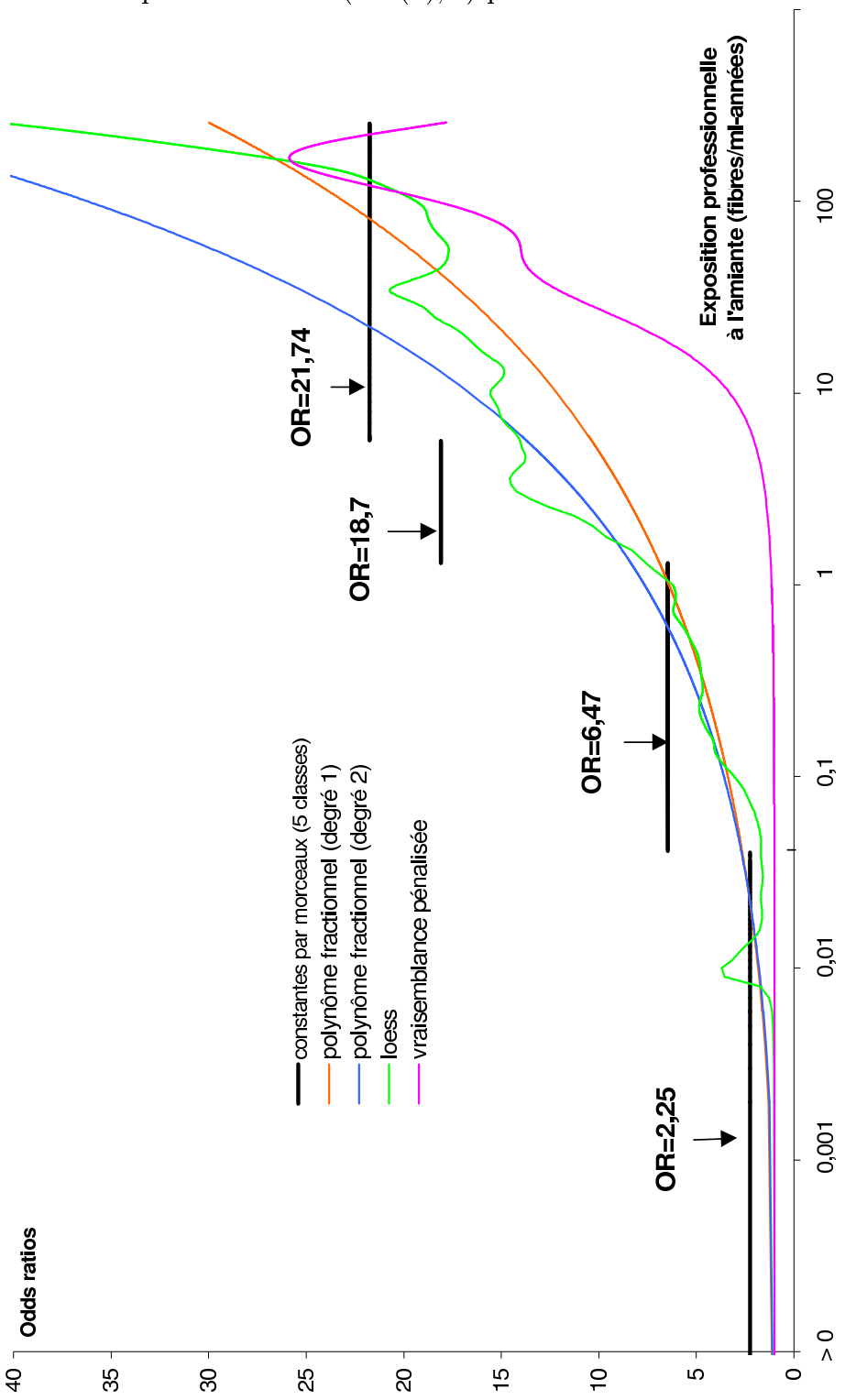


FIG. 3.2 – Représentation de $(OR(x), x)$ pour les meilleurs estimateurs



3.7 Sélection de modèles par la méthode de Birgé-Massart

Nous cherchons à approcher la fonction f par une fonction constante par morceaux comme dans la section 3.5.1. Pour pouvoir bien approcher la fonction f, nous allons essayer toutes les partitions possibles et retenir la meilleure grâce à la méthode de sélection de modèles de Birgé-Massart [2] proposée dans le cadre des processus linéaires gaussiens. Cette approche a notamment été employée par Letué [10] pour l'ajustement d'une fonction de régression par des histogrammes dans le modèle de Cox.

3.7.1 Principe de la méthode

La méthode consiste à choisir une collection de modèles supposés avoir de bonnes propriétés d'approximation pour la fonction à estimer. Tout d'abord, on estime la fonction f sur chacun des modèles en minimisant un contraste. On obtient ainsi une collection d'estimateurs. Puis on sélectionne un estimateur dans la collection en minimisant sur tous les modèles un critère issu uniquement des données, qui est la somme du contraste pris en l'estimateur et d'un terme de pénalité. Ce terme de pénalité represente la complexité de la collection de modèles et est proportionnel au nombre de paramètres à estimer divisé par le nombre d'observations. Il est nécessaire de définir une pénalité convenable pour que le modèle sélectionnée réalise un compromis entre l'erreur d'approximation et l'erreur d'estimation.

3.7.2 Modèle

Soit m une partition de dimension D_m de la variable d'exposition d'étendue $[x_{(1)} : x_{(n)}]$. m est définie de la façon suivante :

$$m = \bigcup_{j=1}^{D_m} I_j$$

telle que

$$\begin{cases} I_k =]\tau_{k-1}, \tau_k] \quad k = 2, \ldots, D_m \text{ et } I_1 = [\tau_0, \tau_1] \text{ avec } \tau_0 = 0 \text{ et } \tau_{D_m} = x_{(n)} \\ I_k \bigcap I_{k'} = \emptyset \text{ pour } k = 1, \ldots, D_m \text{ et } k' = 1, \ldots, D_m \end{cases}$$

On note $\mathcal{S}_m$ le **modèle** associé à la partition m, comme étant le sous-espace linéaire des fonctions constantes par morceaux construites sur la partition m :

$$S_m = \left\{ u = \sum_{k=1}^{D_m} u_k \mathbb{I}_{I_k}, (u_k)_{k=1,\ldots,D_m} \in \mathbb{R}^{D_m} \right\}$$

3.7.3 Construction de l'estimateur par minimum de contraste

Nous choisissons comme contraste empirique l'opposé de la log-vraisemblance définie pour toute fonction $u \in \mathcal{S}_m$ par :

$$\begin{aligned} \gamma_n(u) &= -\frac{1}{n} \sum_{i=1}^{n} [y_i \log p(x_i) + (1 - y_i) \log \{1 - p(x_i)\}] \\ &= -\frac{1}{n} \sum_{i=1}^{n} \left[y_i \log \frac{e^{u(x_i)}}{1 + e^{u(x_i)}} + (1 - y_i) \log \frac{1}{1 + e^{u(x_i)}} \right] \end{aligned}$$

Le maximum de vraisemblance de l'estimateur $\widehat{f}_m$ de la fonction f sur le modèle $\mathcal{S}_m$ est la fonction unique dans $\mathcal{S}_m$ qui minimise $\gamma_n(u)$ pour tout $u \in \mathcal{S}_m$:

$$\hat{f}_m = \underset{u \in \mathcal{S}_m}{\operatorname{argmin}} \gamma_n(u).$$

On montre facilement que le contraste empirique est minimal pour $u_k = \bar{y}_k$, $k = 1, \ldots, D_m$ où $\bar{y}_k$ est la moyenne empirique de la variable réponse y sur le segment I_k. Ainsi l'estimateur du mimimum de contraste empirique de f sur le sous espace linéaire $\mathcal{S}_m$ est :

$$\widehat{f}_m = \sum_{k=1}^{D_m} \bar{y}_k \mathbb{I}_{I_k}.$$

3.7.4 Collection de modèles et d'estimateurs

On note $\mathcal{M}_n$ l'ensemble de toutes les partitions possibles construites sur la grille $\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$ où $x_{(\cdot)}$ représente les valeurs ordonnées de l'exposition. Pour chaque partition m de $\mathcal{M}_n$, nous définissons $\mathcal{S}_m$ le sous espace linéaire des fonctions constantes par morceaux construites sur la partition m :

$$\{\mathcal{S}_m, m \in \mathcal{M}_n\}$$

Ainsi nous définissons la collection d'estimateurs de minimum de contraste :

$$\{\widehat{f}_m, m \in \mathcal{M}_n\}.$$

L'objectif est de sélectionner le meilleur estimateur parmi cette collection. Pour cela, nous utilisons un critère pénalisé ne dépendant que des données.

3.7.5 Critère pénalisé : $crit_n(m) = \gamma_n(\hat{f}_m) + pen_n(m)$

Dans le cadre d'estimation de densité par des histogrammes, Castellan [4] propose un critère pénalisé $crit_n(m) = \gamma_n(\hat{f}_m) + pen_n(m)$ pour sélectionner le meilleur histogramme. Le maximum de vraisemblance est utilisé comme contraste empirique. Dans ce contexte, Castellan, s'appuyant sur les travaux de Birgé-Massart [2], propose comme forme de pénalité :

$$pen_n(m) = \frac{D_m}{n}\left(c_1 \log\frac{n}{D_m} + c_2\right) \quad \text{pour tout } m \in \mathcal{M}$$

où c_1 et c_2 sont des constantes positives.
Le premier terme du critère $crit_n(m)$ est lié à l'ajustement aux observations (plus la dimension de la partition est élevée plus cet ajustement est bon). Le second terme $pen_n(m)$, dépendant de la partition m que par sa dimension D_m, permet de contrôler la dimension de la partition à sélectionner. Le terme $\log\frac{n}{D_m}$ est issu de la complexité de la collection de

partitions $\mathcal{M}_n$. Finalement, cette pénalité est fonction de deux constantes c_1 et c_2. Leurs valeurs sont obtenues par une étude de simulations. En s'inspirant des travaux de Birgé et Rozenholc [3], Castellan trouve par simulation dans le cas d'estimation de densité que $c_1 = 1$ et $c_2 = 2.5$.

Dans le cadre de nos données, nous choisissons une pénalitée proportionnelle à celle proposée par Castellan. Le critère devient donc :

$$crit_n(m) = \gamma_n(\hat{f}_m) + pen_n(m)$$

avec

$$pen_n(m) = \beta \frac{D_m}{n} \left(\log \frac{n}{D_m} + 2.5 \right)$$

où β est la constante de proportionnalité. On peut noter que dans un cas d'estimation d'une fonction de régression (hypothèse gaussienne) par des constantes par morceaux, Lebarbier [9] utilise une pénalité identique.

La pénalité ne depend plus que d'une constante β. Une méthode heuristique est proposée par Birgé et Massart [1] pour déterminer ce facteur.

Comme $pen_n(m)$ ne dépend que de D_m, sélectionner le meilleur estimateur parmi la collection $\{\widehat{f}_m, m \in \mathcal{M}_n\}$ revient à choisir le meilleur parmi la collection $\{\widehat{f}_{\hat{m}_D}, D = 1, \ldots, n\}$, où $\hat{m}_D$ est la meilleure partition de dimension D, définie par

$$\hat{m}_D = \underset{m \in \mathcal{M}_n, |m| = D}{\operatorname{argmin}} \gamma_n(\hat{f}_m)$$

Enfin, la meilleure dimension est définie par :

$$\hat{D} = \underset{D \geq 1}{\operatorname{argmin}} \left[\gamma_n(\hat{f}_{\hat{m}_D}) + \beta \frac{D}{n} \left(\log \frac{n}{D} + 2.5 \right) \right]$$

On notera par la suite $\widehat{f}_D$ pour désigner $\widehat{f}_{\hat{m}_D}$. Il reste donc à déterminer β pour trouver la meilleur dimension D. Notre critère devient donc après cette première étape :

$$crit_n(D) = \gamma_n(\hat{f}_D) + pen_n(D)$$

avec $pen_n(D) = \beta g_n(D)$ et $g_n(D) = \frac{D}{n}\left(\log\frac{n}{D} + 2.5\right)$. En reparamétrisant par $\beta = 2\alpha$, le critère devient :

$$crit_n(D) = \gamma_n(\hat{f}_D) + 2\alpha g_n(D).$$

Il est nécessaire d'obtenir une bonne valeur α pour pénaliser correctement. Nous utilisons la méthode proposée par Birgé-Massart [1] qui montre que pour des "grandes dimensions" $\gamma_n(\hat{f}_D)$ est une fonction affine de $g_n(D)$. La pente $-\hat{\alpha}$ de la courbe représentant la fonction affine est un estimateur de $-\alpha$. Pour une justification en détail se reporter à [9]. Finalement,

$$crit_n(D) = \gamma_n(\hat{f}_D) + 2\hat{\alpha} g_n(D) \tag{3.2}$$

Dans certaines situations, l'estimation de la pente pose quelques problèmes (liés à la détermination de "grande dimension"). Pour résoudre ce problème, une autre heuristique, basée sur la définition de la pénalité minimale, est développée par Birgé et Massart [1]. La pénalité minimale est la grandeur $\hat{\alpha} g_n(D)$. On cherche la pénalité optimale qui est le double de cette pénalité minimale (3.2). La pénalité minimale représente la plus petite pénalité avec laquelle la partition sélectionnée est de dimension raisonnable. Cela signifie que lorsque le seuil de la pénalité minimale est atteint, la dimension de l'estimateur associé devrait chuter subitement. Ainsi, pour estimer la pénalité minimale, il suffit de faire varier la constante de pénalité α appelée "température", de sélectionner la dimension à partir de chaque pénalité par le critère $crit_n^{minimal}(D) = \gamma_n(\hat{f}_D) + \alpha g_n(D)$ et de repérer le passage à la pénalité minimale. La pénalité optimale est obtenue en prenant le double de la pénalité trouvée. Finalement,

$$crit_n(D) = \gamma_n(\hat{f}_D) + 2\hat{\alpha}\frac{D}{n}\left(\log\frac{n}{D} + 2.5\right)$$

et l'estimateur final est $\tilde{f} = \widehat{f}_{\hat{D}}$ où $\hat{D}$ minimise $crit_n(D)$ en D.

3.7.6 Résultats

Nous effectuons deux modélisations différentes. La premiere modélisation correspond à celle effectuée au sous-chapitre 3.5.1. Seules les partitions m où $I_1 = [\tau_0, \tau_1]$ et $\tau_1 = 0$ sont considérées. Les individus non-exposés sont pris comme classe de référence. En pratique, la méthode de sélection est effectuée sur l'échantillon $\tilde{\mathcal{W}}_{n_1}^1$ (individu exposé à l'amiante). Cette modélisation est appelée " modèle avec intercept". Dans la seconde modélisation aucune hypothèse n'est faite sur la classe de référence. Toutes les partitions sont prises en compte et "modèle complet" désignera cette modélisation.

3.7.7 Modèle avec intercept : étude sur l'échantillon $\tilde{\mathcal{W}}_{n_1}^1$

Nous représentons les resultats des deux heuristiques pour trouver la meilleure dimension $\hat{D}$. La figure 3.3 représente la méthode de la pente.

Nous trouvons un coefficient directeur $-\hat{\alpha} = -0.44$. Il permet d'en déduire la pénalité

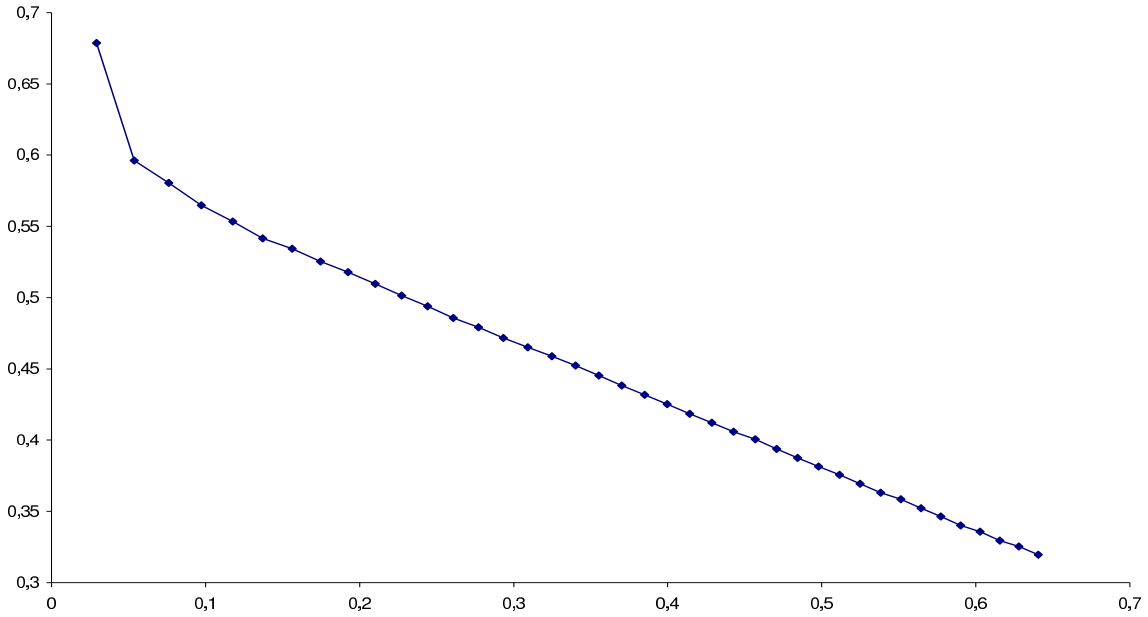


FIG. 3.3 – Représentation de $(g_n(D), \gamma_n(\hat{f}_D))$, pour $D = 1, \ldots, 40$: modèle avec intercept

optimale $2\hat{\alpha}g_n(D)$ et par la suite d'obtenir la dimension de la meilleure partition ; nous trouvons $\hat{D} = 2$. La figure 3.4 représente la méthode de la pénalité minimale. La plus grande chute de dimension a lieu pour $\hat{\alpha}_{min}$=0.459 . La meilleure dimension est aussi $\hat{D} = 2$. En rajoutant le paramètre correspondant à $I_1 = [0, 0]$ (individus non-exposé), le meilleur modèle est de dimension 3. La figure (3.7) représente les résultats de ce modèle en terme de probabilités et d'odds ratio.

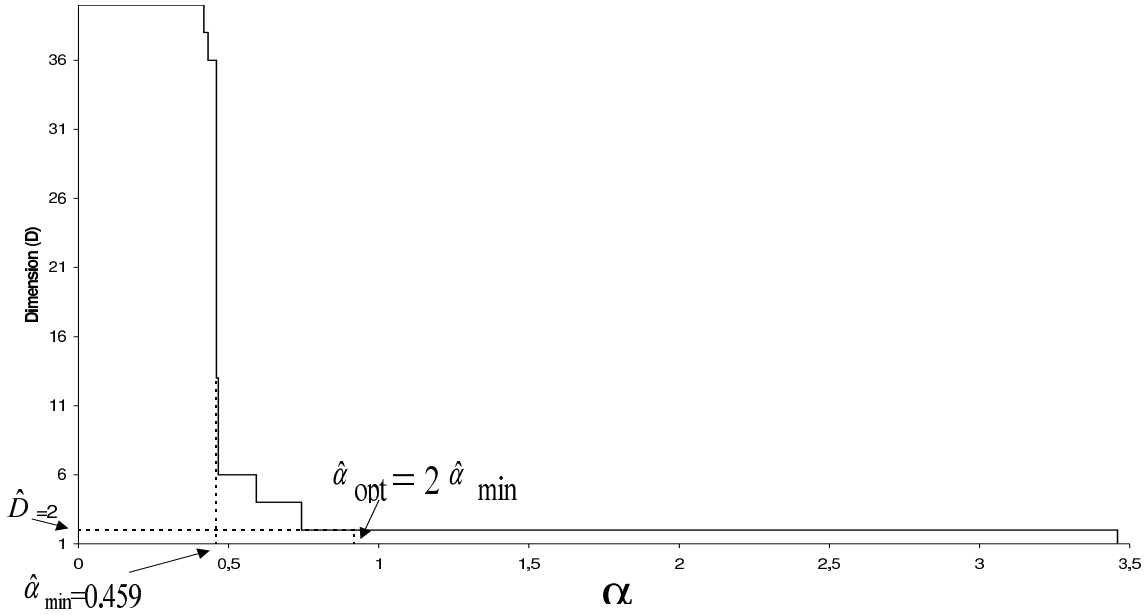


FIG. 3.4 – Recherche de la pénalité minimale ; D fonction de α suivant $crit_n^{minimal}(D)$: modèle avec intercept

3.7.8 Modèle complet

Pour cette modélisation, la pente vaut $-\hat{\alpha} = -0.40$ (voire figure 3.5). Ainsi la pénalité optimale vaut $2\hat{\alpha}g_n(D) = 0.80g_n(D)$ et la dimension correspondante vaut $\hat{D}$=3 . Ce résultat est confirmé par la méthode de la pénalité minimale (voir figure 3.6).
Une représentation graphique de ce modèle est présentée avec le modèle avec intercept sur la figure 3.7.

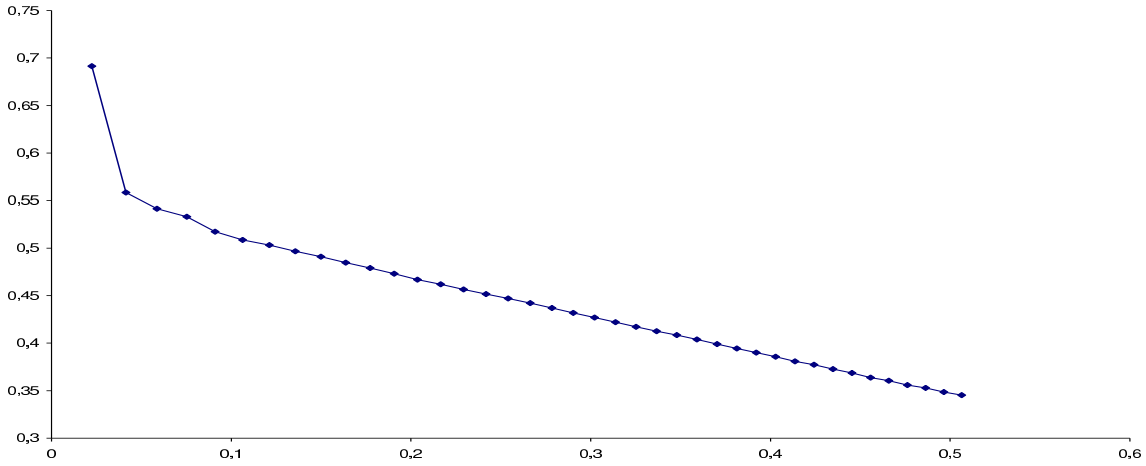


FIG. 3.5 – Représentation de $(g_n(D), \gamma_n(\hat{f}_D))$, pour $D = 1, \ldots, 40$: modèle complet

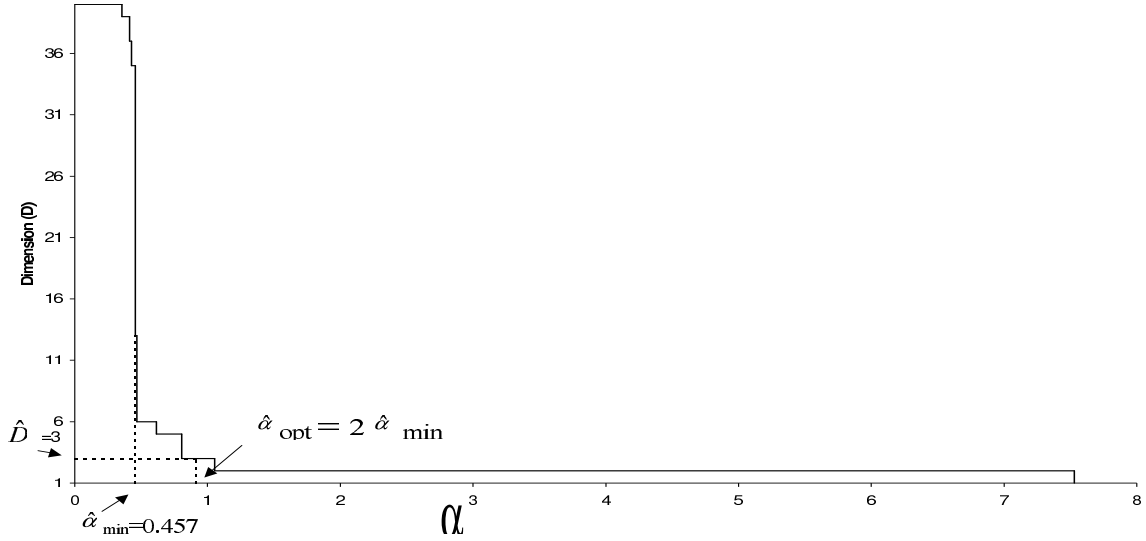
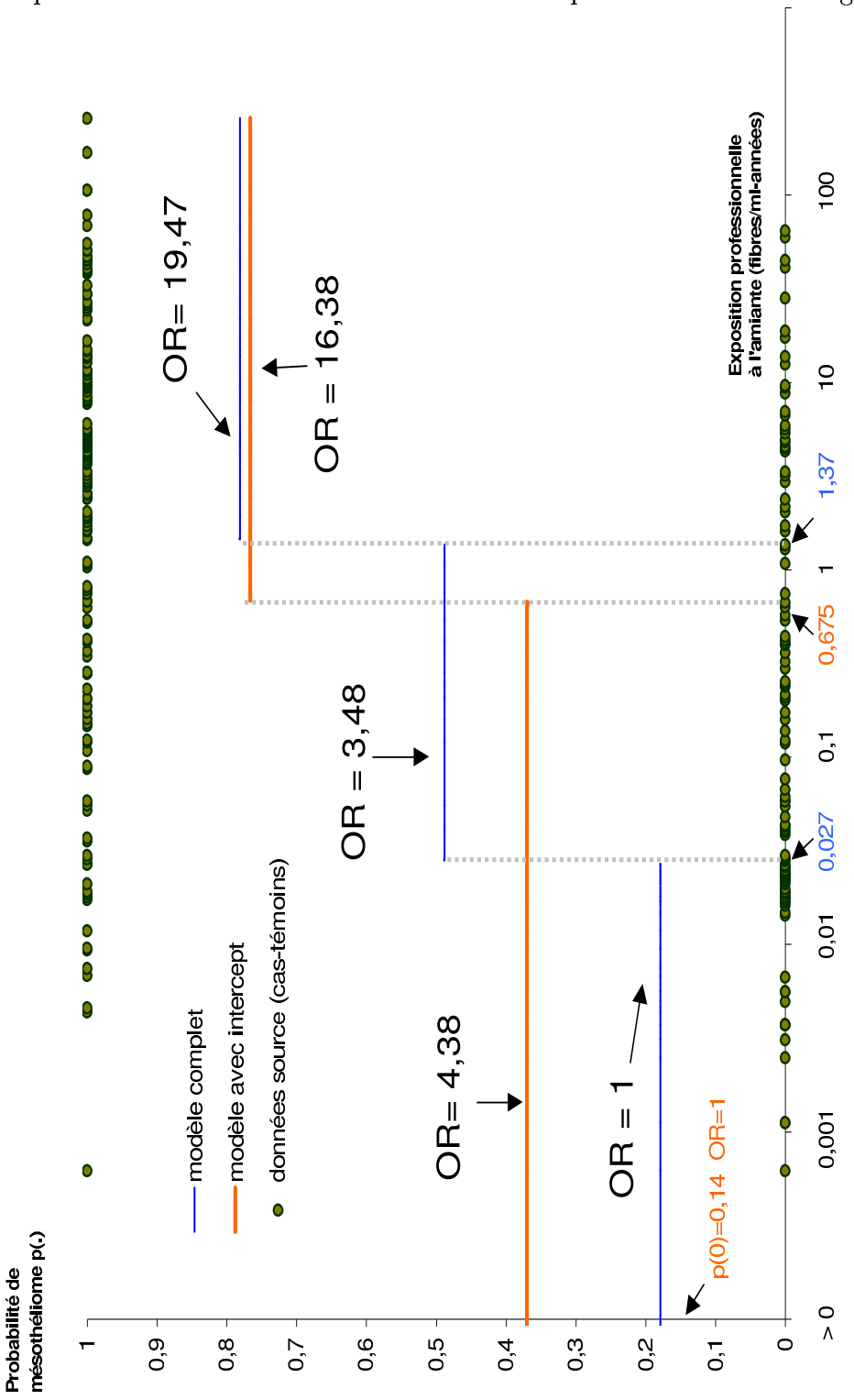


FIG. 3.6 – Recherche de la pénalité minimale ; D fonction de α suivant $crit_n^{minimal}(D)$: modèle complet

FIG. 3.7 – Représentation des deux modèles sélectionnés par la méthode de Birgé-Massart



3.8 Conclusion

La méthode de sélection présentée dans cette section, nous a permis d'essayer dans un premier temps toutes les partitions de la même forme que celle proposée dans le sous-chapitre 3.5.1. Nous avons aussi pu étudier toutes les partitions possibles. Il est clair que le modèle appelé modèle avec intercept est un cas particulier du modèle complet. Finalement le meilleur estimateur est construit sur une partition de dimension 3 avec $I_1 = [0, 0.027],\ I_2 =]0.027, 1.37],\ I_3 =]1.37, 258.3]$. Le risque de mésothéliome est de 3.48 (respectivement 19.47) pour les individus appartenant à I_2 (respectivement I_3) par rapport aux individus appartenant à I_1 (classe de référence).

La comparaison de cette méthode de sélection à celle utilisée par le EIC parait difficile. Le critère EIC a permis de sélectionner le meilleur estimateur parmi des familles d'estimateurs indexées par un seul hyper-paramètre. En effet, la famille d'estimateurs par morceaux définie en 3.5.1 est indexée par k, la dimension du modèle. Pour les polynômes fractionnels, nous avons considéré un nombre fini d'estimateurs (k représente alors le rang de l'estimateur considéré). Dans le cas d'estimateurs non-paramétriques (loess et vraisemblance pénalisée), les familles d'estimateurs sont indexées par le paramètre de lissage. Ainsi toutes les familles sont indexées par un seul hyper-paramètre. Il parait donc raisonable de comparer la valeur maximale de l'EIC des différentes familles. En revanche, la méthode de sélection de modèles proposée par Birgé-Massart définit un estimateur (famille à un estimateur). Il serait risqué de comparer la valeur du EIC de cet estimateur aux valeurs maximales de l'EIC des autres familles. Ce problème de comparaison de familles de modèles de complexités différentes sera discuté dans les perspectives.

Bibliographie

[1] L. Birgé and P. Massart. *A generalized C_p criterion for gaussian model selection.* Tech. rep., Publication Université Paris-VI, 2001.

[2] L. Birgé and P. Massart. J. Eur. Math. Soc. *Gaussian model selection*, 3 :203–268, 2001.

[3] L. Birgé and Y. Rozenholc. *How many bins should be put in a regular histogram.* Tech. rep., Publication Université Paris-VI, 1999.

[4] G. Castellan. *Selection d'histogrammes ou de modeles exponentiels de polynomes par morceaux a l'aide d'un critere de type akaike.* PhD thesis, Université Paris Sud, Orsay, 2000.

[5] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368) :829–836, 1979.

[6] Hosmer D.W. and Lemeshow S. *Applied Logistic Regression.* Wiley and Sons, New York, 1989.

[7] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models.* Chapman and Hall, London, 1990.

[8] Inserm. *Effets sur la santé des principaux types d'exposition à l'amiante.* Éditions INSERM, Expertises Collectives, Paris, 1997.

[9] E. Lebarbier. *Quelques approches pour la détection de ruptures à horizon fini.* PhD thesis, Université de Paris-Sud, 2002.

[10] F. Letué. *Modèle de Cox : estimation par sélection de modèle et modèle de chocs bivarié.* PhD thesis, Université de Paris-Sud, 2000.

[11] B. Liquet, C. Sakarovitch, and D. Commenges. Bootstrap choice of estimators in non-parametric families : an extension of EIC. *Biometrics*, 2002 (in press).

[12] P. Royston and D. G. Altman. Regression using fractional polynomials of continuous covariates : parsimonious parametric modelling (with discussion). *Applied Statistics*, 43 :429–467, 1994.

Chapitre 4

Choix d'estimateurs semi-paramétriques en présence de données incomplètes

4.1 Introduction

Dans le chapitre 2, nous avons proposé un critère permettant de choisir un estimateur parmi une famille d'estimateurs paramétriques ou une famille d'estimateurs semi-paramétriques. Ce critère est basé sur une approximation de la distance de Kullback-Leibler. Nous proposons d'étendre ce point de vue à des données incomplètes, ce que l'on peut rencontrer en analyse de survie. Dans certaines situations, il est important de pouvoir estimer de façon lisse la fonction de risque λ. Cette fonction est utilisée fréquemment en analyse de données de survie. On note $\widehat{\lambda}_h^{\mathcal{W}}(\cdot)$ un estimateur lisse de $\lambda(\cdot)$ avec h représentant le paramètre de lissage et $\mathcal{W}$ représentant l'échantillon de n variables aléatoires i.i.d. Notre objectif est de choisir par un critère d'information le paramètre de lissage pour une famille d'estimateurs et aussi de choisir entre différentes familles d'estimateurs.

Dans un cadre de données complètes, le terme pertinent de l'information de Kullback-Leibler (KL) est l'espérance de la log vraisemblance d'une nouvelle observation. En présence

de données censurées, il est difficile d'estimer ce critère. Nous proposons alors comme critère possible le CELL qui est l'espérance conditionnelle de la log-vraisemblance observée d'un nouvel échantillon qui est une copie de l'échantillon observé :

$$\mathrm{CELL}\{\widehat{\lambda}_h^{\mathcal{W}}\} = \frac{1}{n}\mathrm{E}\left\{\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}')|\mathcal{W}\right\}$$

ou $\mathcal{W}'$ a la même distribution que $\mathcal{W}$ ($\mathcal{W}' \stackrel{d}{=} \mathcal{W}$) et $\mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}')$ est fonction de vraisemblance de l'estimateur $\widehat{\lambda}_h^{\mathcal{W}}(\cdot)$ pour les observations $\mathcal{W}'$. Dans le même esprit qu'Akaike qui considère l'espérance de l'information de Kullback-Leibler, nous considérons comme second critère l'espérance de CELL :

$$\mathrm{ELL}(\widehat{\lambda}_h) = \mathrm{E}\left\{\mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}})\right\} = \frac{1}{n}\mathrm{E}\left\{\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}')\right\}$$

Dans la mesure où les deux critères sont calculables, nous disposons de deux procédures pour sélectioner h. La procédure adaptative qui dépend de $\mathcal{W}$

$$h_{\mathrm{CELL}}(\mathcal{W}) = \underset{h}{\mathrm{argmax}}\ \mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}})$$

et la procédure non-adaptative (en accord avec les propositions d'Akaike)

$$h_{\mathrm{ELL}} = \underset{h}{\mathrm{argmax}}\ \mathrm{ELL}(\widehat{\lambda}_h^{\mathcal{W}})$$

Afin de comparer les deux procédures de sélection, on note que :

$$\max_{h'} \mathrm{CELL}(\widehat{\lambda}_{h'}^{\mathcal{W}}) \quad \geq \quad \mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}}),\ \forall\ h$$

En appliquant l'espérance aux deux membres de l'inégalité, nous obtenons :

$$\mathrm{E}\left\{\max_{h'}\ \mathrm{CELL}(\widehat{\lambda}_{h'}^{\mathcal{W}})\right\} \quad \geq \quad \mathrm{E}\left\{\mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}})\right\},\ \forall\ h$$

et donc, nous avons

$$\begin{aligned} \mathrm{E}\left\{\max_h \mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}})\right\} &\geq \max_h \mathrm{E}\left\{\mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}})\right\} \\ \mathrm{ELL}(\widehat{\lambda}_{h_{\mathrm{CELL}}}^{\mathcal{W}}) = \mathrm{E}\left\{\mathrm{CELL}(\widehat{\lambda}_{h_{\mathrm{CELL}}}^{\mathcal{W}})\right\} &\geq \mathrm{ELL}(\widehat{\lambda}_{h_{\mathrm{ELL}}}^{\mathcal{W}}) \end{aligned}$$

Ainsi la méthode adaptative peut théoriquement avoir des meilleures performances que la méthode non-adaptative, jugée sur le critère ELL. Mais il n'est pas évident que la supériorité puisse être exploitée par des critères pratiques. Des simulations (non présentées) ont montré qu'il était illusoire d'estimer $\mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}})$. Ainsi le critère que nous proposons d'étudier est le ELL. Plusieurs estimateurs de ELL seront considérés et utilisés pour comparer des modèles stratifiés et des modèles à risques proportionnels.

Dans la suite de ce chapitre, nous présentons notre article (soumis à la publication) dans "Biometrics", intitulé "Estimating the expectation of the log likelihood with incomplete data for choosing an estimator in semi-parametric families".

Estimating the expectation of the log-likelihood with incomplete data for choosing an estimator in semi-parametric families

B. LIQUET and D.COMMENGES

INSERM U330, ISPED

146 rue Léo Saignat

33076 Bordeaux cedex, France

Summary

A criterion for choosing an estimator in a family of semi-parametric estimators from incomplete data is proposed. This criterion is the expected observed log-likelihood (ELL). Adapted versions of this criterion in case of censored data and in presence of explanatory variables are exhibited. We show that likelihood cross-validation is an estimator of ELL and we exhibit three bootstrap estimators. A simulation study considering both families of kernel and penalized likelihood estimators of the hazard function (indexed on a smoothing parameter) demonstrates good results of LCV and a bootstrap estimator called ELL_{bboot}. When using penalized likelihood an approximated version of LCV also performs very well. The use of these estimators of ELL is exemplified on the more complex problem of choosing between stratified and unstratified proportional hazards models. An example is given for modeling the effect of sex and educational level on the risk of developing dementia.

KEY WORDS : bootstrap, cross-validation, Kullback-Leibler information, proportional hazard model, semi-parametric, smoothing

4.2 Introduction

The problem of model choice is obiously one of the most important in statistics. Probably one of the first solution to a model choice problem was given by Mallows (1973) who proposed a criterion (C_p) for selecting explanatory variables in linear regression problems. This problem of selection of variables was studied by many authors in more general regression models (Copas, 1983). The celebrated Akaike criterion (Akaike, 1974) solved the problem of parametric model selection. This criterion called AIC (An Information Criterion) was based on an approximation of the Kullback-Leibler distance (Akaike, 1973). Criterions improving AIC for small samples have been proposed : AIC_c (Hurvich and Tsai, 1989) and EIC which is a bootstrap estimation (Ishiguro et al., 1997). Finally in the case of missing data, Cavanaugh and Shumway (1998) proposed a variant of AIC. A closely related, but more difficult problem, is that of choice of a smoothing parameter in smoothed semi-(or non-) parametric estimation of functions. These functions may be density function (Silverman, 1986), effect functions of an explanatory variable (Hastie and Tibshirani, 1990) or hazard functions (O'Sullivan, 1998, Joly et al., 1998). Smoothing methods are in particular kernel smoothing methods and penalized likelihood. In simple regression problems, versions of AIC and AIC_c are available (Hurvich et al., 1998) and simple versions of the cross-validation criterion have been proposed : CV, GCV (Green and Silverman, 1994). However in general problems only the likelihood cross-validation criterion (LCV) (O'Sullivan, 1998) and bootstrap techniques, in particular, extension of EIC (Liquet et al., 2003) are available. In some problems approximations of the mean integrated square error (MISE) are available (Ramlau-Hansen, 1983).

Liquet et al. (2003) have introduced a general point of view which is to choose an estimator among parametric or semi-parametric families of estimators according to a criterion which is an approximation of the Kullback-Leibler distance; they have shown on some simulation studies that the best criterions were EIC and LCV. They treated a general multivariate regression problem. The aim of this paper is to extend this point of view to the case where incomplete data are observed. The data may be incomplete because of

right or interval-censoring for instance. This is not a trivial extension : indeed, it becomes clear that for using relatively simply the bootstrap approach, the theoretical criterion to be estimated must be changed. The proposed criterion is the expectation of the (observed) log-likelihood (ELL) rather than the Kullback-Leibler distance.

We define the ELL criterion in section 4.3 and give useful versions of it for use with right-censored data and with explanatory variables (where partial and conditional likelihood respectively are used). In section 4.4 we exhibit three bootstrap estimators of ELL and show that LCV also estimates ELL. Section 4.5 presents simulation studies for comparing the four estimators together with the Ramlau-Hansen approach for hazard functions using kernel smoothing methods or penalized likelihood.

In section 4.6, we show an application of these criterions to a more complex problem, which is to compare stratified and unstratified proportional hazards models. Our particular application is modeling onset of dementia as a function of sex and education level (coded as a binary variable). We could consider a proportional hazard for both variables or stratified models on one variable, or making four strata. No method has been proposed to our knowledge to compare such different semi-parametric models. We propose to compare them using the ELL criterion, in practice using LCV or a bootstrap estimator, and apply these methods to the data of the PAQUID study, a large cohort study on dementia (Letenneur et al., 1994).

4.3 The expected log-likelihood as theoritical criterion

4.3.1 Definitions and notations

First we consider $\mathcal{W} = (W_1, \ldots, W_n)$ a sample of i.i.d. variables with values in R^d and common distribution $F_W(\cdot)$. We will consider that W_i brings information on a random variable T_i, the distribution of which we want to estimate. We will develop the theory in this general framework ; then we shall specialize to the case of right-censored observations

where $W_i = (\tilde{T}_i, \delta_i)$; $\tilde{T}_i = \min(T_i, C_i)$ and $\delta_i = \mathrm{I}_{[T_i \leq C_i]}$. T_i is the time of the event of interest and C_i is a censoring variable ; T_i and C_i are assumed to be independently and identically distributed (i.i.d.) with distribution $F(\cdot)$ for T_i and $F_C(\cdot)$ for C_i. In the sequel, we denote by f and f_C the probability density functions, S and S_C the survival functions of T and C respectively. In survival analysis, one of the most interpretable function is the hazard function, $\lambda(t) = \frac{f(t)}{S(t)}$; any of the function λ, f, S determines the distribution. We denote by $\widehat{\lambda}_h^{\mathcal{W}}(\cdot)$ a family of estimators of $\lambda(\cdot)$, where h most often represents a smoothing parameter. To any particular estimator $\widehat{\lambda}_h^{\mathcal{W}}(\cdot)$ corresponds an estimator $\widehat{f}_h^{\mathcal{W}}(\cdot)$. Our aim is to propose an information criterion to choose the smoothing parameter for a family of estimators and also to choose between different families of estimators.

4.3.2 The expected log-likelihood

For uncensored data, the useful part of the Kullback-Leibler information criterion, measuring the distance between $\widehat{f}_h^{\mathcal{T}}(\cdot)$ and f, is the conditional expectation of the log-likelihood of a future observation $T^{'}$ given $\mathcal{T}$

$$\mathrm{KL}(\mathcal{T}) = \mathrm{E}\left\{\log \widehat{f}_h^{\mathcal{T}}(T^{'})|\mathcal{T}\right\} \tag{4.1}$$

where $\mathcal{T} = (T_1, \ldots, T_n)$ and $T^{'}$ an additional observation having the distribution F and being independent of the sample $\mathcal{T}$. In presence of incomplete data, it is difficult to estimate this criterion because $\mathcal{T}$ is not observed : it is not possible to directly estimate the expectation in (4.1) by bootstrap. Instead, we define a first criterion as the conditional expectation of the observed log-likelihood of a new sample which is a copy of the original sample, given the original sample :

$$\mathrm{CELL}\{\widehat{\lambda}_h^{\mathcal{W}}\} = \frac{1}{n}\mathrm{E}\left\{\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}^{'})|\mathcal{W}\right\} \tag{4.2}$$

where $\mathcal{W}^{'}$ has the same distribution as $\mathcal{W}$ (we will use in the following the notation $\mathcal{W}^{'} \stackrel{d}{=} \mathcal{W}$) and $\mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}^{'})$ is the likelihood function of the estimator $\widehat{\lambda}_h^{\mathcal{W}}(\cdot)$ for observation

$\mathcal{W}'$. In a sense this is an observed information criterion.

In the special case of uncensored observations ($\mathcal{W} = \mathcal{T} = (T_1, \ldots, T_n)$), $\mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}})$ is equivalent to the criterion $\mathrm{KL}(\mathcal{T})$. Indeed we have

$$\begin{aligned}
\mathrm{CELL}\{\widehat{\lambda}_h^{\mathcal{W}}\} &= \frac{1}{n}\mathrm{E}\left\{\log\prod_{i=1}^{n}\widehat{f}_h^{\mathcal{W}}(W_i^{'})|\mathcal{W}\right\} \\
&= \frac{1}{n}\mathrm{E}\left\{\sum_{i=1}^{n}\log\widehat{f}_h^{\mathcal{W}}(W_i^{'})|\mathcal{W}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left\{\log\widehat{f}_h^{\mathcal{W}}(W_i^{'})|\mathcal{W}\right\}
\end{aligned}$$

As $W_i^{'}$ are i.i.d., we have

$$\mathrm{CELL}\{\widehat{\lambda}_h^{\mathcal{W}}\} = \mathrm{E}\left\{\log\widehat{f}_h^{\mathcal{W}}(W_i^{'})|\mathcal{W}\right\} = \mathrm{KL}(\mathcal{T})$$

It is interesting to discuss the relation between CELL and the Akaike approach (Akaike, 1973). $\mathrm{CELL}\{\widehat{\lambda}_h^{\mathcal{W}}\}$ depends on $\mathcal{W}$ because the expectation is taken conditionally on $\mathcal{W}$ whereas Akaike proposed to consider the expectation of the Kullback-Leibler information : $\mathrm{E}\{\mathrm{KL}(\mathcal{W})\}$. In the extended framework we can in the same spirit consider the expectation of CELL : the second criterion is defined by

$$\mathrm{ELL}(\widehat{\lambda}_h) = \mathrm{E}\left\{\mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}})\right\} = \frac{1}{n}\mathrm{E}\left\{\log\mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}^{'})\right\}. \tag{4.3}$$

This criterion does not depend on $\mathcal{W}$ and judges a procedure of estimation $\widehat{\lambda}_h$ that can be applied to any $\mathcal{W}$ of same distribution. So, supposing that we could compute the two criterions, we have two procedures to select the index parameter h. Some simulations results (not shown here) have made clear that it is illusory to try to estimate $\mathrm{CELL}\{\widehat{\lambda}_h^{\mathcal{W}}\}$. Thus the criterion that we propose is, in accordance with the pinciple of Akaike (see also DeLeeuw, 1992), the non-conditional expectation of the log-likelihood, ELL. Indeed it is relatively easy to show that for a parametric model, AIC defined as $-2\log\mathcal{L}+2p$ (p being the number of parameters) is an estimator of ELL (see Cavanaugh and Shumway, 1998).

4.3.3 Case of right-censored data

In presence of right-censored data as defined in section 4.3.1, the likelihood $\mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}')$ is :

$$\mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}') = \prod_{i=1}^{n} \{\widehat{f}_h^{\mathcal{W}}(\tilde{T}_i')\}^{\delta_i'} \{\widehat{S}_h^{\mathcal{W}}(\tilde{T}_i')\}^{1-\delta_i'} \{f_C(\tilde{T}_i')\}^{1-\delta_i'} \{S_C(\tilde{T}_i')\}^{\delta_i'}$$

where $\widehat{f}_h^{\mathcal{W}}(\cdot)$ and $\widehat{S}_h^{\mathcal{W}}(\cdot)$, the estimators of f and S are deduced from $\widehat{\lambda}_h^{\mathcal{W}}(\cdot)$.
The criterion defined in (4.3) can be decomposed in two parts :

$$\mathrm{ELL}(\widehat{\lambda}_h) = \mathrm{ELL}^p(\widehat{\lambda}_h) + \mathrm{E}\{\phi(f_C, S_C, \mathcal{W}')\} \tag{4.4}$$

The first term is defined by :

$$\mathrm{ELL}^p(\widehat{\lambda}_h) = \frac{1}{n}\mathrm{E}\left\{\log \mathcal{L}_p^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}')\right\} \tag{4.5}$$

where

$$\mathcal{L}_p^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}') = \prod_{i=1}^{n} \{\widehat{f}_h^{\mathcal{W}}(\tilde{T}_i')\}^{\delta_i'} \{\widehat{S}_h^{\mathcal{W}}(\tilde{T}_i')\}^{1-\delta_i'}$$

is the partial likelihood (in the sense of Andersen et al., 1993). The second term in (4.4) does not depend on $\widehat{\lambda}_h$; thus maximizing ELL is equivalent to maximize ELL^p. Finally our criterion is :

$$\mathrm{ELL}^p(\widehat{\lambda}_h) = \frac{1}{n}\mathrm{E}\left\{\log \mathcal{L}_p^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}')\right\}$$

that is the ELL criterion applied to the partial likelihood ; this is very fortunate because this avoids estimating the distribution of the censoring variable. Note however that the ELL criterion cannot be applied to the Cox partial likelihood, at least directly : we need a smooth estimate of the hazard function to apply our criterion. Any non-smooth estimate has a value $-\infty$ and is rejected.

4.3.4 Case of explanatory variable

We consider the case of presence of explanatory variables. We note $W_i = (T_i, X_i)$ with T_i the survival time and X_i a vector of covariates for the ith individual. It is assumed that T_i has conditional density function $f(\cdot|x_i)$ given $X_i = x_i$. Our aim is to estimate $\lambda(\cdot|\cdot)$ the corresponding conditional hazard function. We note this estimator $\widehat{\lambda}_h^{\mathcal{W}}(\cdot|\cdot)$ and $\widehat{f}_h^{\mathcal{W}}(\cdot|\cdot)$ the corresponding density. The likelihood $\mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}')$ is :

$$\mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}') = \prod_{i=1}^{n} \{\widehat{f}_h^{\mathcal{W}}(T_i'|X_i')\}\{f_X(X_i')\}$$

where $\widehat{f}_X(\cdot)$ is the marginal density of X_i

With the same reasoning as in 4.3.3, the criterion in (4.3) can be decomposed in two parts :

$$\mathrm{ELL}(\widehat{\lambda}_h) = \mathrm{ELL}^c(\widehat{\lambda}_h) + \mathrm{E}\{\phi(f_X, \mathbf{X}')\} \tag{4.6}$$

where $\mathbf{X}' = (X_1', \ldots, X_n')$. The first term is defined by :

$$\mathrm{ELL}^c(\widehat{\lambda}_h) = \frac{1}{n}\mathrm{E}\left\{\log \mathcal{L}_c^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}')\right\}$$

where

$$\mathcal{L}_c^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}') = \prod_{i=1}^{n} \{\widehat{f}_h^{\mathcal{W}}(T_i'|X_i')\}$$

is the conditionnal likelihood. The second term of (4.6) does not depend on $\widehat{\lambda}_h$; thus maximizing ELL is equivalent to maximizing ELL^c. Finally our criterion is :

$$\mathrm{ELL}^c = \frac{1}{n}\mathrm{E}\left\{\log \mathcal{L}_c^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}')\right\}$$

that is the ELL criterion applied to the conditionnal likelihood ; this is very fortunate because this avoids estimating the distribution of the explanatory variable. Both tricks can be applied when there are both explanatory variables and censoring.

4.4 Estimators of ELL as pratical choice criterions

4.4.1 LCV

LCV is a well known method for model choice. We argue that a theoretical basis of LCV is that it is an estimator of ELL. The method of likelihood cross-validation is a natural development of the idea of using likelihood to judge the adequacy of fit of a statistical model. Let W' be an additional observation, independent of the others. The log-likelihood of the new observations would be $\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(W')$. Since we do not have additional observations, we may omit one of the original observations W_i from the sample used to construct the density estimator (sample $\mathcal{W}^{-i}$), and then use W_i as the observation W', which gives $\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^{-i}}}(W_i)$. Since there is nothing special about the choice of which observation to leave out, the log-likelihood is averaged over all the possible choices of W_i. This leads to the LCV criterion :

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

The expectation of LCV is approximatevely equivalent to ELL ; indeed we have :

$$\begin{aligned} \mathrm{E}\,(\mathrm{LCV}) &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left\{\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^{-i}}}(W_i)\right\} \\ &= \mathrm{E}\left\{\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^{-i}}}(W_i)\right\} = \mathrm{ELL}\{\widehat{\lambda}_h(n-1)\} \end{aligned}$$

where $\widehat{\lambda}_h(n-1)$ is an estimator applied to a sample of size $(n-1)$; for a reasonable family of estimators, we can consider that for large n the estimator is nearly the same when we take n rather than $(n-1)$. So,

$$\mathrm{E}\,(\mathrm{LCV}) \simeq \mathrm{ELL}(\widehat{\lambda}_h).$$

An idea of the order of magnitude of the variance of LCV can be suggested by a formal analogy between LCV and $\widehat{\mathrm{ELL}}(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \log \mathcal{L}^{\lambda}(W_i)$ for fixed λ. Neglecting the differences between the $\widehat{\lambda}_h(\mathcal{W}^{-i})$ and the fact that they depend on $\mathcal{W}$, we may compute $\frac{1}{n}\sum_{i=1}^{n} \frac{\{\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^{-i}}}(W_i) - \mathrm{LCV}\}^2}{n-1}$ as an approximation of the variance of LCV. This formula gives the good order of magnitude in the simulation of section 4.5.1 .

4.4.2 Direct bootstrap method for estimating ELL (ELL_{boot} and ELL_{iboot})

We can directly estimate by bootstrap the expectation of the log-likelihood (ELL). We define this bootstrap estimator as

$$\mathrm{ELL}_{boot} = \frac{1}{n}\mathrm{E}_* \left\{ \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(\mathcal{W}'^*) \right\}$$

where $\mathcal{W}^* = (W_1^*, \ldots, W_n^*)$, $W_j^* \sim \widehat{F}_W$, $\mathcal{W}'^* = (W_1'^*, \ldots, W_n'^*)$ and $W_j'^* \sim \widehat{F}_W$, $\widehat{F}_W$ being the empirical distribution of W_i based on $\mathcal{W}$. We use the notation E_* to remind that the expectation is taken relatively to the estimated distribution $\widehat{F}_W$. In practice, the expectation is approximated by a mean of B repetitions of bootstrap samples ($\mathcal{W}^j \stackrel{d}{=} \mathcal{W}'^j \stackrel{d}{=} \mathcal{W}^*$)

$$\mathrm{ELL}_{boot} \simeq \frac{1}{n}\frac{1}{B}\sum_{j=1}^{B} \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^j}}(W'^j)$$

To improve this criterion, we can iterate the bootstrap method. We define this new estimator as :

$$\mathrm{ELL}_{iboot} = \frac{1}{n}\mathrm{E}_{**} \left\{ \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^{**}}}(\mathcal{W}'^{**}) \right\}$$

where $\mathcal{W}^{**} = (W_1^{**}, \ldots, W_n^{**})$, $W_j^{**} \sim \widehat{F}_{W^*}$, $\mathcal{W}'^{**} = (W_1'^{**}, \ldots, W_n'^{**})$ and $W_j'^{**} \sim \widehat{F}_{W^*}$, $\widehat{F}_{W^*}$ being the empirical distribution of W_i^* based on $\mathcal{W}^*$. E_{**} is calculated with respect to the distribution $\widehat{F}_{W^*}$. The expectation is also approximated by a mean of B repetitions

of bootstrap samples ($\mathcal{W}^j \stackrel{d}{=} \mathcal{W}'^j \stackrel{d}{=} \mathcal{W}^{**}$)

$$\mathrm{ELL}_{iboot} \simeq \frac{1}{n}\frac{1}{B}\sum_{j=1}^{B} \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^j}}(W'^j)$$

4.4.3 Bias corrected bootstrap estimators

To construct this estimator, we first propose a bootstrap estimator of the CELL criterion (which happens to be the log-likelihood itself) and then correct it by estimating its bias. This approach is similar to that used for deriving the EIC (Liquet et al., 2003) criterion available for complete data.

Let $\mathcal{W} = (W_1, \ldots, W_n)$ a sample of i.i.d. variables with $W_i \in R^d$ and common distribution $F_W(\cdot)$. So the mirror sample $\mathcal{W}'$ used in the ELL criterion (4.2) is also a sample of i.i.d. variables W_i'. Thus we have :

$$\mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}') = \prod_{i=1}^{n} \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(W_i')$$

and then the criterion reduces to

$$\mathrm{CELL}\{\widehat{\lambda}_h^{\mathcal{W}}\} \quad = \quad \mathrm{E}\left\{\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(W_i')|\mathcal{W}\right\}$$

A natural estimator is found by replacing F_W (which is the distribution of W_i') with its empirical distribution $\widehat{F}_W$:

$$\widehat{\mathrm{CELL}}(\widehat{\lambda}_h^{\mathcal{W}}) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(W_i') = \frac{1}{n}\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}) \tag{4.7}$$

(that we may recognize as the log-likelihood of $\mathcal{W}$ for the distribution specified by $\widehat{\lambda}_h^{\mathcal{W}}(\cdot)$) However this generally overestimates $\mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}})$ because $\mathcal{W}$ has already been used to determine $\widehat{\lambda}_h^{\mathcal{W}}(\cdot)$ for fixed h. We propose to estimate the bias

$$b = \mathrm{E}\left\{\widehat{\mathrm{CELL}}(\widehat{\lambda}_h^{\mathcal{W}}) - \mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}})\right\} \tag{4.8}$$

by bootstrap. With $\mathcal{W}^* = (W_1^*, \ldots, W_n^*)$ and $W_j^* \sim \widehat{F}_W$, we find

$$\widehat{b}(\mathcal{W}) \quad = \quad \mathrm{E}_* \left\{ \widehat{\mathrm{CELL}}(\widehat{\lambda}_h^{\mathcal{W}^*}) - \mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}^*}) | \mathcal{W} \right\} \tag{4.9}$$

Replacing $\mathcal{W}$ by $\mathcal{W}^*$ in (4.7) and (4.2), we find :

$$\widehat{b}(\mathcal{W}) = \frac{1}{n} \mathrm{E}_* \left[\log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(\mathcal{W}^*) - \mathrm{E}_* \left\{ \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(\mathcal{W}^{*'}) | \mathcal{W}^* \right\} \middle| \mathcal{W} \right] \tag{4.10}$$

where $\mathcal{W}'^* \stackrel{d}{=} \mathcal{W}^*$ and $\mathcal{W}'^*$ is conditionally independent of $\mathcal{W}^*$ given $\mathcal{W}$. Because the $W_i'^*$ are i.i.d. and have the distribution $\widehat{F}_W$, we have for the second term of (4.10) :

$$\begin{aligned} \mathrm{CELL}(\widehat{\lambda}_h^{\mathcal{W}^*}) = \mathrm{E}_* \left\{ \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(\mathcal{W}'^*) | \mathcal{W}^* \right\} \quad &= \quad n \mathrm{E}_* \left\{ \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(W_i'^*) | \mathcal{W}^* \right\} \\ &= \quad \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(\mathcal{W}) \end{aligned}$$

So we have :

$$\widehat{b}(\mathcal{W}) \quad = \quad \frac{1}{n} \mathrm{E}_* \left\{ \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(\mathcal{W}^*) - \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^*}}(\mathcal{W}) | \mathcal{W} \right\}.$$

The expectation conditional on $\mathcal{W}$ can be approximated by a mean of B evaluations with (bootstrap samples) $\mathcal{W}^j$ taken at random from the distribution of $\mathcal{W}^*$:

$$\widehat{b}(\mathcal{W}) \simeq \frac{1}{n} \frac{1}{B} \sum_{j=1}^{B} \left\{ \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^j}}(\mathcal{W}^j) - \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}^j}}(\mathcal{W}) \right\}.$$

Finally our corrected estimator of CELL is :

$$\mathrm{ELL}_{bboot} = \frac{1}{n} \log \mathcal{L}^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}) - \widehat{b}(\mathcal{W}). \tag{4.11}$$

Note that the bias b in (4.8) approximated by $\widehat{b}(\mathcal{W})$ does not depend on $\mathcal{W}$ and is in fact the bias of the log-likelihood considered as an estimator of ELL ($b = \mathrm{E}\{\widehat{\mathrm{CELL}}(\widehat{\lambda}_h^{\mathcal{W}})\} - \mathrm{ELL}$). So the bias corrected criterion actually estimates ELL, hence its name.

Remark : for all the bootstrap methods when treating right-censored observations the bootstap expectations have to be conditionned on having at least one uncensored observation because the estimator are not defined otherwise.

4.5 Simulation

We have compared ELL_{boot}, ELL_{iboot}, ELL_{bboot} and LCV using both families of kernel and penalized likelihood estimators of hazard functions. We have included the Ramlau-Hansen method when using kernels, a popular method for estimating hazard functions (Andersen et al., 1993) and the approximated LCV popular for estimating hazard functions using penalized likelihood (O'Sullivan, 1998 ; Joly et al., 1998). We compare the criteria when using kernel smoothing in 4.5.1 and penalized likelihood in 4.5.2.

4.5.1 Kernel estimator

The smoothed Nelson-Aalen estimator is

$$\widehat{\lambda}(t) = \frac{1}{h} \int K\left(\frac{t-u}{h}\right) d\widehat{A}(u)$$

where $K(\cdot)$ is a kernel function, $\widehat{A}(\cdot)$ is the Nelson-Aalen estimator of $A(\cdot)$, the cumulative hazard function, and h is the bandwith parameter. Ramlau-Hansen (1983) has proposed an estimator of the MISE (mean integrated square error) based on an approximated cross-validation method for estimating h ; we call it the RH method. We apply gaussian kernels to allow the use of the different criteria. Indeed, if we used a kernel with compact support, we risk for small h to have LCV criteria equal to $-\infty$. For the criteria based on bootstrap, kernels with compact support are prohibited since the bootstrap expectations are theoretically equal to $-\infty$ for bandwidth lower than the range of the observed event times. We consider problems where the density near zero is very low so there is no edge effect near zero.

The data were generated from a mixture of gamma distributions. We generated random

samples $T_1, ..., T_n$ of i.i.d. failure times and $C_1, ..., C_n$ of i.i.d. censoring times; the C_i were independent of the T_i. So the observed samples were $(\tilde{T}_1, \delta_1), ..., (\tilde{T}_n, \delta_n)$ where $\tilde{T}_1 = \min(T_i, C_i)$ and $\delta_i = \mathrm{I}_{[T_i \leq C_i]}$. The density of T was a a mixture of Gamma $\{0.4\Gamma(t; 40, 1) + 0.6\Gamma(t; 80, 1)\}$, with the probability density functions $\Gamma(t; \alpha, \gamma) = \frac{\alpha^\gamma t^{\gamma-1} e^{-\alpha t}}{\Gamma(\gamma)}$. The probability density function of C_i was a simple Gamma : $\Gamma(t; 90, 1)$, $\Gamma(t; 90, 1.1)$ and $\Gamma(t; 90, 1.3)$ corresponding to a percentage of censoring around 15%, 25% and 50% respectively. Samples of sizes 30, 50 and 100 were generated. Figure 4.1 displays the smoothed Nelson-Aalen estimate chosen by ELL_{bboot} and the true hazard function for one simulated example from a mixture of gamma.

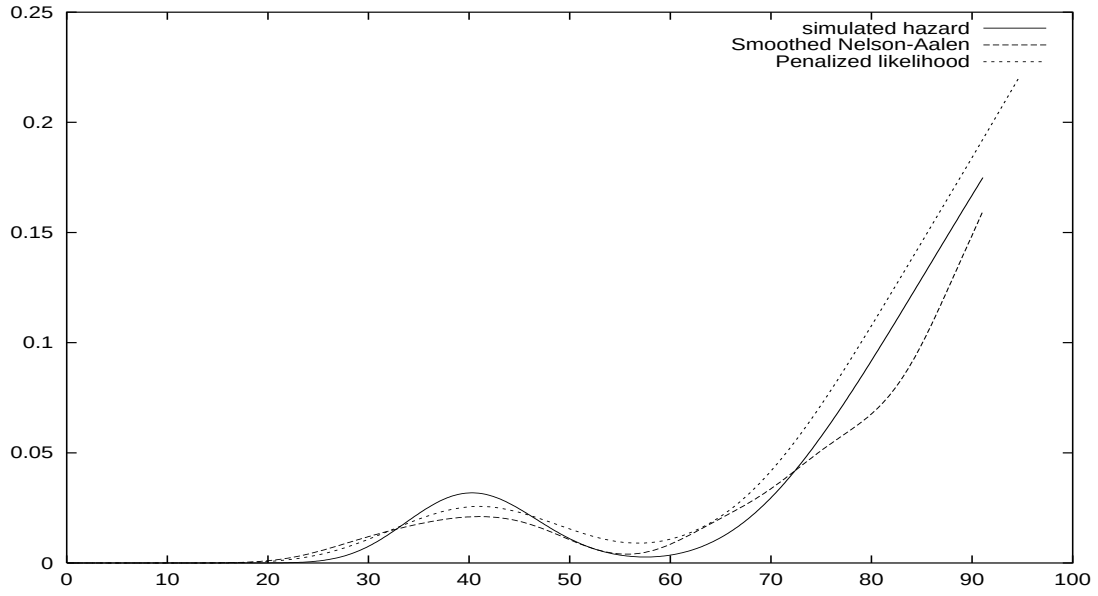


FIG. 4.1 – *True hazard function, smoothed Nelson-Aalen estimator and penalized likelihood estimate chosen by* ELL_{bboot} *for a simulated example. The sample size is 50, with* 15% *right-censored observations.*

Each simulation involved 100 replications. For each replication we computed the useful part of the Kullback-Leibler information (KL) between the true density function f and the estimators chosen by each criterion

$$\mathrm{KL}(f; \widehat{f}_h^{\mathcal{W}}) = \int_J \log \widehat{f}_h^{\mathcal{W}}(t) f(t) dt$$

where $J =]0; T_{max}]$. We do not take T_{max} equal to $+\infty$, because for large times t when there is censoring, we do not have enough information to determine $\widehat{f}_h^{\mathcal{W}}(t)$. T_{max} was chosen for each simulation such as

$$\Pr\{\mathrm{E}(n_{T_{max}}) \geq 1\} = 0.95$$

where $n_{T_{max}}$ represents the risk set at time T_{max}. We computed, for each simulation presented, the average of KL and its standard error. Since KL generally takes negative values we give in tables 4.1-4.3 the values of -KL : low values then correspond to estimators close to the true distribution. First we present in table 4.1 the results of the simulation comparing the optimal criterion KL and the new criterion ELL.

The two theoretical criteria give practically the same results. We note some differences

TAB. 4.1 – *Average Kullback-Leibler information* KL *for the kernel estimator for estimating the hazard function of the mixture of gamma* $(0.4\Gamma(t,a,b)+0.6\Gamma(t,c,d))$ *for bandwith chosen by* ELL *and* KL*, based on 100 replications.*

	$-\mathrm{KL}(\lambda_h^{\mathcal{W}})$ for kernel estimators	
n	KL	ELL
15% of censoring		
30	3.96(0.005)	3.96(0.005)
50	3.98(0.003)	3.99(0.003)
100	4.01(0.002)	4.01(0.002)
25% of censoring		
30	3.89(0.004)	3.91(0.005)
50	3.93(0.004)	3.93(0.004)
100	3.95(0.002)	3.95(0.002)
50% of censoring		
30	3.81(0.02)	3.92(0.04)
50	3.80(0.009)	3.84(0.02)
100	3.80(0.005)	3.80(0.005)

only when there is little information (small sample size and high censoring level). The average of -KL obtained for the pratical criteria are given in table 4.2. These averages can

be compared to an optimal value, the value of KL when estimators are chosen using the true ELL.We may note that RH yielded in all cases much higher (worse) values of -KL than the other criteria.

TAB. 4.2 – *Average Kullback-Leibler information* KL *for the kernel estimator for estimating the hazard function of the mixture of gamma* $\{0.4\Gamma(t,a,b)+0.6\Gamma(t,c,d)\}$ *for each criterion based on 100 replications.*

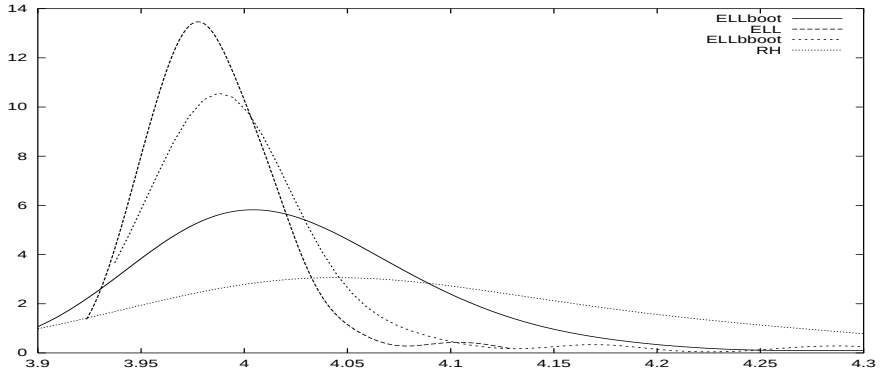
	$-\mathrm{KL}(\lambda_h^{\mathcal{W}})$ for kernel estimators					
n	ELL	ELL_{bboot}	LCV	ELL_{iboot}	ELL_{boot}	RH
15% of censoring						
30	3.96(0.005)	4.00(0.009)	4.01(0.02)	3.98(0.005)	4.04(0.01)	4.19(0.06)
50	3.99(0.003)	4.00(0.006)	4.00(0.008)	4.00(0.005)	4.04(0.01)	4.22(0.06)
100	4.01(0.002)	4.02(0.002)	4.02(0.002)	4.02(0.002)	4.05(0.005)	4.12(0.02)
25% of censoring						
30	3.91(0.005)	3.94(0.009)	3.96(0.01)	3.92(0.006)	3.98(0.01)	4.26(0.08)
50	3.93(0.004)	3.95(0.007)	3.96(0.01)	3.94(0.006)	3.99(0.01)	4.2(0.06)
100	3.95(0.002)	3.96(0.002)	3.96(0.002)	3.96(0.002)	3.99(0.007)	4.10(0.03)
50% of censoring						
30	3.92(0.04)	3.99(0.07)	4.04(0.07)	4.01(0.07)	4.02(0.08)	4.36(0.1)
50	3.84(0.02)	3.85(0.02)	3.91(0.03)	3.85(0.02)	3.88(0.03)	4.18(0.09)
100	3.80(0.005)	3.81(0.005)	3.83(0.02)	3.80(0.005)	3.84(0.008)	3.95(0.03)

The ELL_{boot} criterion, although better than RH, had in pratically all the cases higher values than the other criteria. The differences were very small between LCV, ELL_{iboot} and ELL_{bboot} although for high censoring level and small sample sizes, LCV tended to perform not as well as the bootstrap methods. For all the simulations, the three competitive criteria

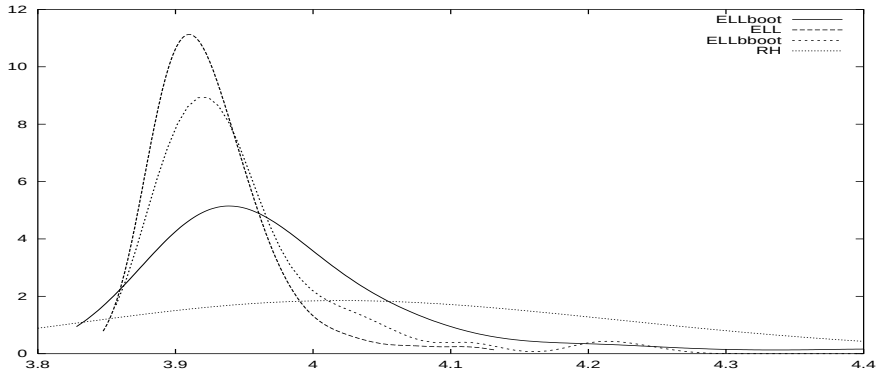
had values of KL quite close to the values given by ELL.

Figure 4.2 displays the p.d.f. of the $-$KL distance for ELL, ELL_{bboot}, ELL_{boot}, and RH for the simulation with $n = 50$ (the other distributions are not presented).

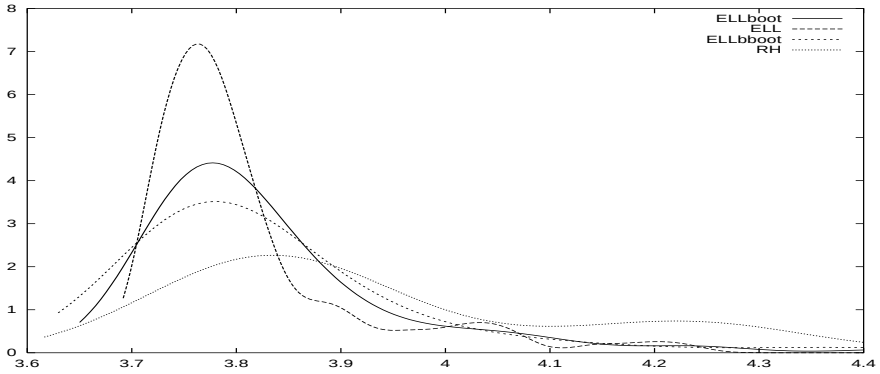
The curves in this figure are themselves kernel density estimates with bandwith parameter selected by ELL_{bboot} method. These curves confirm the superiority of the ELL_{bboot} criterion over ELL_{boot} and RH since the distribution is more concentrated on lower values. Corresponding curves for LCV and ELL_{iboot} were very close to those of ELL_{bboot} and are not shown for lisibility of this figure. For comparing the pratical criteria in their ability of estimating ELL, we represent in figure 4.3a, the expected values of the different criteria and ELL as functions of the bandwith parameter h. These curves are the average curves over 1600 replicatons of curves for simulation with n=50 and 15% of censoring. We represent also in figure 4.3b the variance of each criterion. In this example, for a reasonable bandwith, all the criteria had nearly the same variance around 0.020. Using the approximation proposed in section 4.4.1, we find an empirical variance of LCV around 0.024. We can see that ELL_{boot} overestimates ELL. The other criteria are very close to ELL. However we can observe a slight significant bias. For instance, for a bandwith equals to 4.2 (close to the optimal) the differences between the mean of the different criteria LCV, ELL_{iboot} and ELL_{bboot} and ELL are respectively -0.005, 0.01, 0.02 while the standard errors of theses means are 0.004. LCV seems to have the lowest bias.



(a)

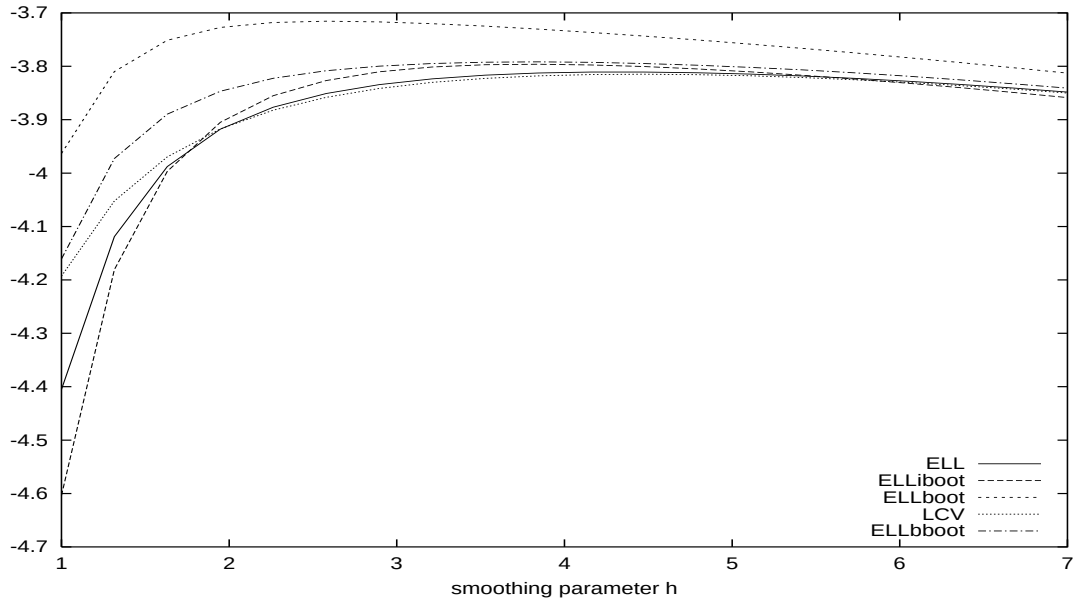


(b)

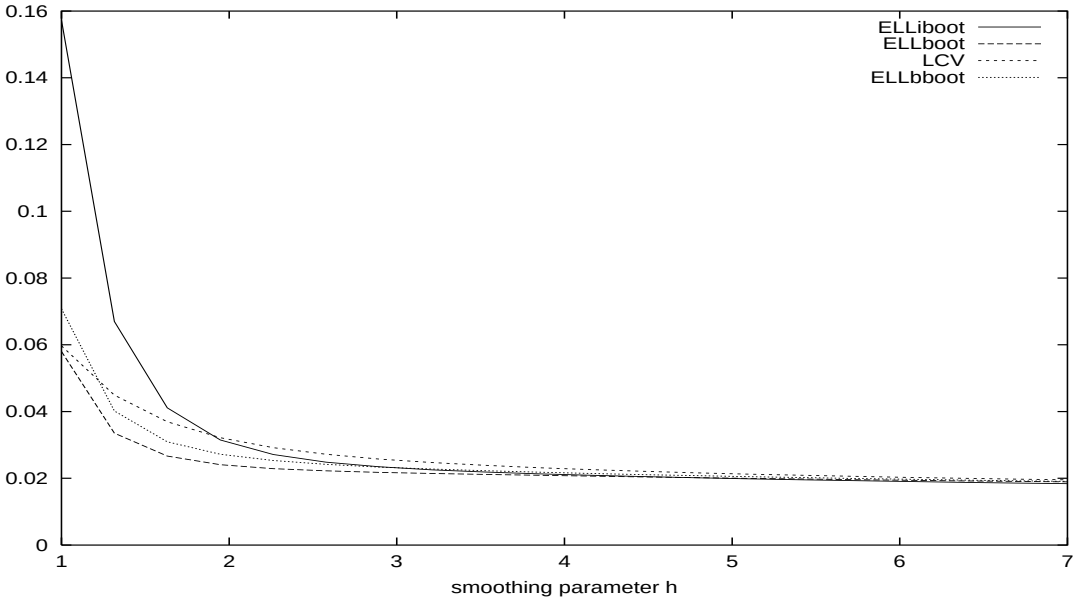


(c)

FIG. 4.2 – *Kernel density estimates of the distribution of* $-\mathrm{KL}$ *for the kernel estimators using* ELL_{bbboot}, ELL_{boot}, RH *and the optimal* KL *: with* $n = 50$ *and (a)* 15% *of censoring, (b)* 25% *of censoring, (c)* 50% *of censoring.*



(a)



(b)

FIG. 4.3 – *Properties of the different criteria for the simulation with* $n = 50$ *and* 15% *of censoring and for 1600 replications : (a) bias (b) variance.*

4.5.2 Penalized likelihood estimator

Another approach to estimate the hazard function is to use penalized likelihood :

$$p\mathcal{L}_h(\mathcal{W}) \quad = \quad \log \mathcal{L}_p^{\lambda}(\mathcal{W}) - h \int \lambda^{''2}(u)du \tag{4.12}$$

where $\mathcal{L}_p^{\lambda^{\mathcal{W}}}$ is the partial log-likelihood (in the sense of section 4.3.3) and h is a positive smoothing parameter which controls the tradeoff between the fit of the data and the smoothness of the function. Maximization of (4.12) over the desired class of functions defines the maximum penalized likelihood estimator (MPLE) $\widehat{\lambda}_h^{\mathcal{W}}$. The solution is then approximated on a basis of splines. Such an approach has been proposed by O'Sullivan (1998) and Joly et al. (1998). The main advantage of the penalized likelihood approach over the kernel smoothing method is that there is no edge problem ; the drawback is that it is more computationally demanding. The method of likelihood cross-validation (LCV) may be used to select h. To circumvent the computational burden of the LCV a one-step Newton-Raphson expansion has been proposed by O'Sullivan (1988) ; we denote this approximation by LCV_a.

ELL_{bboot} and ELL_{iboot} are also applicable to select the smoothing parameter for penalized likelihood estimators. Figure 4.1 displays the penalized likelihood estimate chosen by ELL_{bboot} and the true hazard function for one simulated example. We have compared LCV_a, LCV, ELL_{bboot} and ELL_{iboot} to ELL in a short simulation study (penalized likelihood estimators require more computation than kernel estimators). We used the sample with size $n = 50$, generated in section 4.5.1. The results of the simulation are summarized in table 4.3.

TAB. 4.3 – *Average Kullback-Leibler information* KL *for penalized likelihood estimator for each criterion* ($n = 50$).

	$-\mathrm{KL}(\lambda_h^{\mathcal{W}})$				
% of censoring	ELL	LCV	LCV_a	ELL_{bboot}	ELL_{iboot}
15%	3.99(0.003)	4.00(0.005)	4.00(0.005)	4.00(0.006)	4.06(0.01)
25%	3.93(0.006)	3.98(0.006)	3.99(0.009)	4.01(0.02)	4.08(0.02)
50%	3.96(0.008)	4.00(0.02)	4.07(0.03)	4.06(0.03)	4.32(0.06)

For penalized likelihood estimators, the differences were small between LCV, LCV_a and ELL_{bboot} ; ELL_{iboot} seemed to be less satisfactory.

4.6 Choosing between stratified and unstratified survival models

4.6.1 Method

The estimators of ELL can be used to choose between stratified and unstratified survival models. Consider right-censored data as defined in section 4.3.1 and let $\mathbf{X} = (X_1, \ldots, X_n)$ a vector of binary variable (coded 0/1). Finally, we note $\mathcal{W} = (W_1, \ldots, W_n)$ with $W_i = (\tilde{T}_i, \delta_i, X_i)$ the observed data. We propose to use the ELL_{bboot} or the LCV_a criteria, to choose between a proportional hazards model and a stratified model. We define by

$$\lambda(t|X_i) = \lambda^0(t) \exp \beta X_i \quad i = 1, \ldots, n$$

the proportional hazards model (Cox, 1972) and by

$$\lambda(t|X_i) = \begin{cases} \lambda^0(t) \text{ if } X_i = 0 \\ \lambda^1(t) \text{ if } X_i = 1 \end{cases}$$

the stratified model. To estimate these two models, we may use the penalized likelihood approach. In the proportional hazards regression model, $\widehat{\lambda}_h^0(\cdot)$ and $\widehat{\beta}$ maximize the penalized log-likelihood :

$$p\mathcal{L}_h(\mathcal{W}) \;=\; \log \mathcal{L}_p^{\lambda^0,\beta}(\mathcal{W}) - h\int {\lambda^0}''^2(u)du$$

In the stratified model, $\widehat{\lambda}_h^0(\cdot)$ and $\widehat{\lambda}_h^1(\cdot)$ maximize :

$$\begin{aligned} p\mathcal{L}_h(\mathcal{W}) &= \log \mathcal{L}_p^{\lambda^0,\lambda^1}(\mathcal{W}) - h\int \left\{ {\lambda^0}''^2(u) + {\lambda^1}''^2(u) \right\} du \\ &= \log \mathcal{L}_p^{\lambda^0}(\mathcal{W}^0) - h\int {\lambda^0}''^2(u)du + \log \mathcal{L}_p^{\lambda^1}(\mathcal{W}^1) - h\int {\lambda^1}''^2(u)du \end{aligned}$$

where $\mathcal{W}^0 = (W_1^0, \ldots, W_{n_0}^0)$ with $W_i^0 = (\tilde{T}_i, \delta_i, X_i = 0)$ and $\mathcal{W}^1 = (W_1^1, \ldots, W_{n_1}^1)$ with $W_i^1 = (\tilde{T}_i, \delta_i, X_i = 1)$. We can remark that, we do not estimate separately $\lambda^0(\cdot)$ and $\lambda^1(\cdot)$ on the sample $\mathcal{W}^0$ and $\mathcal{W}^1$. $\lambda^0(\cdot)$ and $\lambda^1(\cdot)$ are estimated using the same smoothing parameter ; thus the family of estimators $\widehat{\lambda}_h(\cdot|\cdot)$ of the proportional hazards model and the family of estimator $\widehat{\lambda}_h(\cdot|\cdot)$ of the stratified model have both just one hyper-parameter h. Therefore, we can discriminate between these two models (we return on this theoretical issue in the discussion). The LCV_a criterion could be applied to select h in the two models and thus to choose between them. It is appealing to apply in addition to the condition of the remark of section 4.4.3, the stronger condition $\sum X_i^{'} = n_1$. This has the advantage on conditioning on an ancillary statistic (the sample sizes in the strata, which does not carry information) and to yield the addition formula (4.13) below. The conditional criterion is thus :

$$\mathrm{ELL}_c(\widehat{\lambda}_h) = \mathrm{E}\left[\log \mathcal{L}_p^{\widehat{\lambda}_h^{\mathcal{W}}}(\mathcal{W}^{'}) | \sum X_i^{'} = n_1\right].$$

where $\mathcal{W}^{'} \stackrel{d}{=} \mathcal{W}$,$\mathcal{W}^{'} = (W_1^{'}, \ldots, W_n^{'})$ with $W_i^{'} = (\tilde{T}_i^{'}, \delta_i^{'}, X_i^{'})$.

To calculate $\mathrm{ELL}_c(\widehat{\lambda}_h)$ we use ELL_{bboot} defined in (4.11) with each bootstrap sample j

that satisfies the condition $\sum_{i=1}^{n} X_i^j = n_1$. For the stratified estimator, we note that :

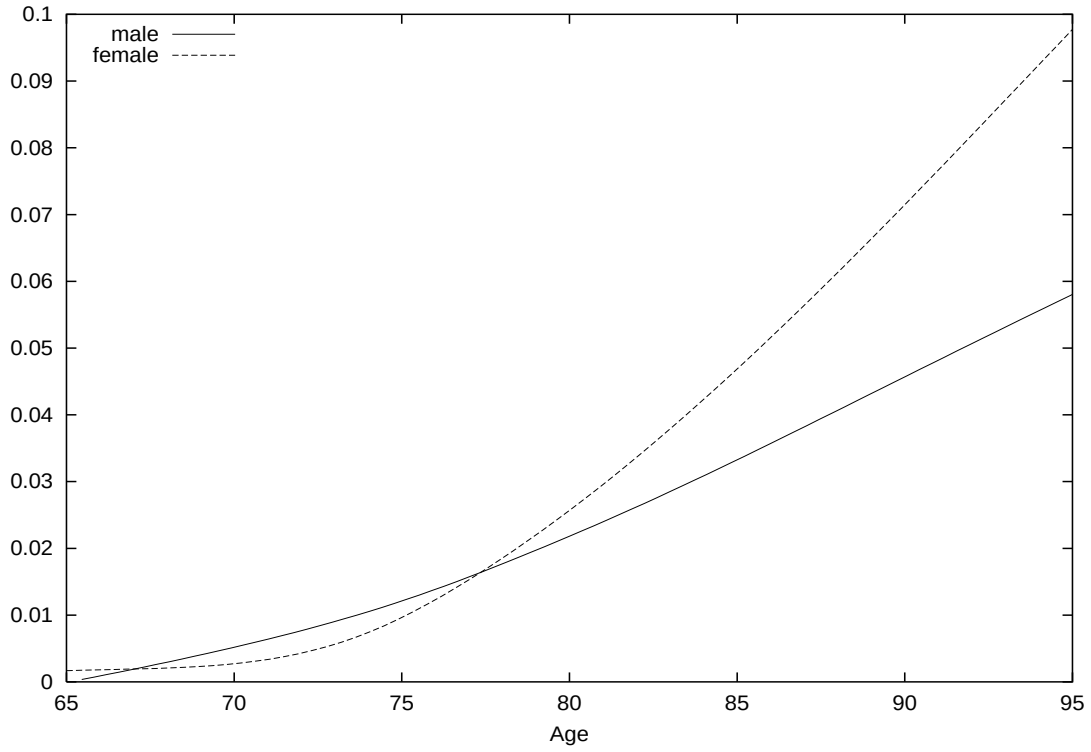
$$\mathrm{ELL}_c(\widehat{\lambda}_h) = \mathrm{ELL}(\widehat{\lambda}_h^0) + \mathrm{ELL}(\widehat{\lambda}_h^1) \tag{4.13}$$

So, in practice for each h we estimated $\mathrm{ELL}(\widehat{\lambda}_h^0)$ and $\mathrm{ELL}(\widehat{\lambda}_h^1)$ by (4.11) applied separately to $\mathcal{W}^0$ and $\mathcal{W}^1$ then computed $\mathrm{ELL}_c(\widehat{\lambda}_h)$ by (4.13). To minimize the different selection criteria we use a golden section search.

4.6.2 Example

We analysed data from the Paquid study (Letenneur et al., 1994), a prospective cohort study of mental and physical aging that evaluates social environment and healh status. The Paquid study is based on a large cohort randomly selected in a population of subjects aged 65 years or more, living at home in two departments of southwest France (Gironde and Dordogne). There were 3675 non demented subjects at entry in the cohort and each subject has been visited six times or less, between 1988 and 2000 ; 431 incident cases of dementia were observed during the folow up. The risk of developing dementia was modeled as a function of age. As prevalent cases of dementia were excluded, data were left-truncated and the truncation variable was the age at entry in the cohort (for more details see Commenges et al., 1998). Two explanatory variables were considered : sex (noted S) and educational level (noted E). In the sample, there are 2133 women and 1542 men. Educational level was classified into two categories : no primary school diploma and primary school diploma (Letenneur et al., 1999). The pattern of observations involved interval censoring and left truncation. It is straightfoward to extent the theory described above to that case. We use the likelihood for interval censoring and left truncated data defined in Commenges (2002). For the sake of simplicity, we kept here the survival data framework, treating death as censoring rather than the more adapted multistate framework. We were first interested in the effect of sex. The penalized likelihood estimate was used to compare the risk of dementia for men and women with a stratified model (model A) (figure 4.4) using ELL_{bboot} for choosing the smoothing parameter.

FIG. 4.4 – *Estimates of the hazard function of dementia for male (solid line) and female (dotted line) chosen by* ELL_{bboot} *criterion.*



The penalized likelihood estimate using the LCV_a criterion was very close to the one obtained with ELL_{bboot}. It appears that women tend to have a lower risk of dementia than men before 78 years and a higher risk above that age and shows a non proportional hazard model. Indeed the proportional hazards model (model B) had lower value for both LCV_a and ELL_{bboot} than the stratified model (table 4.4).

Another important risk factor for dementia is educational level. As the proportional hazards assumption does not hold, we performed several analyses on the educational level stratified on sex. We considered three models. The stratified proportional hazards model (model C) :

$$\lambda(t|S_i, E_i) = \begin{cases} \lambda_h^0(t) \exp \beta E_i \text{ if } S_i = 0 \text{ (women)} \\ \lambda_h^1(t) \exp \beta E_i \text{ if } S_i = 1 \text{(men)} \end{cases}$$

the proportional hazard model performed separately (model D) :

$$\lambda(t|S_i, E_i) = \begin{cases} \lambda_h^0(t) \exp \beta_0 E_i \text{ if } S_i = 0 \text{ (women)} \\ \lambda_h^1(t) \exp \beta_1 E_i \text{ if } S_i = 1 \text{(men)} \end{cases}$$

the model stratified on both sex and educational level (model E) :

$$\lambda(t|S_i, E_i) = \begin{cases} \lambda_h^{0,0}(t) \text{ if } S_i = 0 \text{ and } E_i = 0 \\ \lambda_h^{1,0}(t) \text{ if } S_i = 1 \text{ and } E_i = 1 \\ \lambda_h^{0,1}(t) \text{ if } S_i = 0 \text{ and } E_i = 0 \\ \lambda_h^{1,1}(t) \text{ if } S_i = 1 \text{ and } E_i = 1 \end{cases}$$

Table 4.4 presents the results of the different models. The two criteria give the same conclusion : the best model is the stratified proportional hazard model (highest values ; model C).

TAB. 4.4 – *Comparison of the stratified and proportional hazards models according* ELL_{bboot} *and* LCV_a *criterion ; A and B : unstratified and stratified models on sex ; C, D, E : 3 models for educational level stratified on sex.*

	ELL_{bboot}	LCV_a
model A	-0.4124	-0.4129
model B	-0.4128	-0.4134
model C	-0.4062	-0.4072
model D	-0.4064	-0.4074
model E	-0.4069	-0.4077

4.7 Conclusion

We have presented a general criterion for selection of semi-parametric models from incomplete observations. This theoretical criterion, the expectation of the observed log-likelihood (ELL) performs nearly as well as the optimal KL distance (which is very difficult to estimate in this setting) as soon as there is enough information. We have shown that LCV estimates ELL. LCV and two proposed bootstrap estimators yield nearly equivalent results; ELL_{bboot} seems the best bootstrap estimator. The approximate version of LCV (for penalized likelihood) also performs very well and thus appears as the method of choice for this problem, due to the short computation time it requires. When no approximation of LCV is available, bootstrap estimators such as ELL_{bboot} are competitive because the amount of computation can be more flexibly tuned than for LCV.

ELL can be used for choosing a model in semi-parametric families. An important example is the choice beetween stratified and unstratified survival models. We have shown that this could be done using LCV or a bootstrap estimator of ELL in the case where all the models are indexed by a single hyper-parameter. This raises a completely new problem which is how to compare family of models of different complexities, i.e indexed by a different number of hyper-parameters. For instance this problem would arise if we compared a proportionnal hazards model (1 hyper-parameter) to a stratified model with one hyper-parameter for each stratum. We conjecture that there is a principle of parsimony at the level of the hyper-parameter, similar to that known for the ordinary parameters.

Aknowledgements

We thank Pierre Joly for his help in running the penalized likelihood programs and Luc Letenneur and Jean-François Dartigues for allowing the use of the PAQUID data.

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281, Budapest. Akademiai kiado.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Andersen, P. K., Borgan, R., Gill, R. and Keiding, D. (1993). *Statistical models based on counting processes*. Springer-Verlag, New-York.

Cavanaugh, J. E. and Shumway, R. H. (1998). An Akaike information criterion for model selection in the presence of incomplete data. *Journal of Statistical Planning and Inference* **67**, 45–65.

Commenges, D. (2002). Inference for multistate models from interval-censored data. *Statistical Methods in Medical Research* **11**, 1–16.

Commenges, D., Letenneur, L., Joly, P., Alioum, A. and Dartigues, J. (1998). Modelling age-specific risk: application to dementia. *Statistics in Medecine* **17**, 1973–1988.

Copas, J. B. (1983). Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society B* **45**, 311–354.

Cox, D. (1972). Regression models and life tables (with discussion). *Journal Royal Statistical Society B* **34**, 187–220.

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

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models.* Chapman and Hall, London.

Hurvich, C. M., Simonoff, J. and Tsai, C. (1998). Smoothing parameter selection in non-parametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society B* **60**, 271–293.

Hurvich, C. M. and Tsai, C. (1989). Regression and time series model selection in small samples. *Biometrika* **76**, 297–307.

Ishiguro, M., Sakamoto, Y. and Kitagawa, G. (1997). Bootstrapping log likelihood and EIC, an extension of AIC. *Ann. Inst. Statist. Math* **49**, 411–434.

Joly, P., Commenges, D. and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia. *Biometrics* **54**, 185–194.

Letenneur, L., Commenges, D., Dartigues, J. and Barberger-Gateau, P. (1994). Incidence of dementia and alzheimer's disease in elderly community residents of south-western france. *Int. J. Epidemiol.* **23**, 1256–1261.

Letenneur, L., Gilleron, V., Commenges, D., Helmer, C., Orgogozo, J. and Dartigues, J. (1999). Are sex and educational level independent predictors of dementia and alzheimer's disease? incidence data from the paquid project. *J. Neurol. Neurosurg. Psychiatry.* **66**, 177–183.

Liquet, B., Sakarovitch, C. and Commenges, D. (2003). Bootstrap choice of estimators in non-parametric families: an extension of EIC. *Biometrics. In press* .

Mallows, C. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM J. Sci. Stat. Comput.* **9**, 363–379.

Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics* **11**, 453–466.

Silverman, B. (1986). *Density estimation for statistics and data analysis.* Chapman and Hall, London.

Conclusions et Perspectives

Conclusion générale

Tout d'abord, nous avons proposé une méthode et un programme déterminant le niveau de signification pour une série de codages d'une variable explicative dans une régression logistique. Les simulations sur le risque de première espèce et sur la puissance ont mis en évidence l'avantage du calcul exact sur les autres méthodes de correction (méthode de Bonferroni, méthode d'Efron). Ces simulations ont également permis de fournir des stratégies pour le choix et le nombre de codages de la variable explicative. Enfin ce travail a permis de confirmer l'association faite entre le risque de démence et le HDL cholestérol.

Par ailleurs, nous avons proposé un critère général, le EIC, pour le choix d'estimateurs semi-paramétriques. Des simulations sur des modèles de régression ont montré la supériorité de ce critère, bien que la validation croisée appliquée à la vraisemblance (LCV) obtienne des résultats voisins. D'un point de vue pratique, le calcul par bootstrap (le EIC) est plus long que le LCV quand le nombre de répliques est plus important que la taille de l'échantillon. Dans le cas contraire, le EIC est moins coûteux en temps que le LCV.

Ce critère a été utilisé pour déterminer la meilleure modélisation représentant l'effet de l'amiante sur le risque de mésothéliome dans une étude épidémiologique menée par le Laboratoire Santé Travail et Environnement (Université Bordeaux 2). Parmi les différentes modélisations de l'effet de l'amiante sur le risque de mésothéliome, les polynômes fraction-

nels fournissent les meilleurs résultats. L'estimateur du risque par une approche utilisant des constantes par morceaux donne également aussi de bons résultats. Cette modélisation a été étudiée plus précisément en utilisant la méthode de sélection de modèles de Birgé-Massart. Cette méthode a permis de choisir la meilleure partition (nombre de classes et position des points de coupures) qui est d'habitude choisie de façon arbitraire.

Enfin, un critère de sélection général a été défini dans le cadre de données incomplètes. Ce critère théorique est basé sur l'espérance de la log-vraisemblance observée (ELL). Des simulations ont montré que ce critère théorique a des performances très proches de celles du critère optimal, KL (distance de kullback-leibler). Plusieurs critères pratiques permettant d'estimer le ELL sont disponibles. Le LCV et l'estimateur par bootstrap (ELL_{bboot}) obtiennent les meilleurs performances. L'approximation du LCV pour les estimateurs par vraisemblance pénalisée donne aussi de bons résultats et requiert peu de temps de calcul. Lorsque cette approximation n'est pas disponible, l'estimateur par bootstrap devient plus approprié pour régler le temps de calcul.
Le ELL est aussi utilisé pour choisir entre des modèles stratifiés et des modèles à risques proportionnels. Les critères, LCV et le ELL_{bboot}, sont alors employés pour sélectionner un modèle indexé par un seul hyper-paramètre. Cela soulève un nouveau problème qui est de comparer des familles de modèles de complexités differentes.

Perspectives

Dans la continuité de ce travail, plusieurs perspectives sont envisageables. Elles concernent essentiellement le choix de modèles pour des données incomplètes. Le premier développement concerne la validation croisée appliquée à la vraisemblance (LCV). Le LCV est finalement un estimateur de ELL. De plus sa version approchée a de bonnes performances et limite considérablement le temps de calcul dans les approches par vraisemblance pénalisée. Il serait donc intéressant de définir une version approchée pour les estimateurs à noyau.

Dans le chapitre 4, nous avons comparé des modèles à risques proportionnels à des modèles stratifiés. Afin de considérer des familles indexées par le même nombre d'hyper-paramètres, nous avons décidé que les modèles stratifiés, aient le même paramètre de lissage dans chaque strate. Cette restriction peut poser problème dans le cas où les strates ont des tailles disproportionnées. En effet, un paramètre de lissage identique à toutes les strates pourrait être inapproprié pour certaines strates. La solution serait de considérer une autre grandeur que le paramètre de lissage pour indéxer le modèle. Cette grandeur devrait s'apparenter à un degré de liberté et devrait être fonction de l'information disponible dans chaque strate.

Une perspective plus ambitieuse concerne la comparaison de familles de modèles de complexité différente. En effet, il serait intéressant de pouvoir comparer par exemple un modèle à risques proportionnels avec un seul hyper-paramètre (le paramètre de lissage pour la fonction de risque de base) à un modèle stratifié avec un hyper-paramètre pour chaque strate. Cette difficulté a aussi été rencontrée pour comparer la procédure de sélection par la méthode de Birgé-Masart aux familles d'estimateurs indexées par un hyper-paramètre. Ce problème est équivalent a celui rencontré quand la log-vraissemblance maximisée est considérée comme un estimateur de l'espérance de la log vraisemblance. Dans les approches paramétriques, on tient compte de la complexité du modèle en corrigeant ce biais par le nombre de paramètres. D'une manière similaire on pourrait envisager de corriger l'estimation de ELL pour les valeurs des hyper-paramètres qui maximisent $\widehat{\mathrm{ELL}}$, en tenant compte du nombre d'hyper-paramètres.

Annexe A

Le critère AIC

Nous reprenons ici les principales étapes de construction du critère AIC développé par Akaike. Nous nous appuyons sur l'article de Bozdogan [2] où une théorie générale du AIC est proposée.

Dérivation du AIC

Données et Modèles

Nous observons un échantillon $\mathcal{X} = (x_1, \ldots, x_n)$ de R^d. Chaque x_i est une réalisation i.i.d. d'une variable aléatoire X_i de densité $f(\cdot|\boldsymbol{\theta}^*)$ avec $\boldsymbol{\theta}^* \in R^K$. Nous proposons de modéliser cet échantillon par les K modèles suivants :

$$\mathcal{M}_k = \{f(\cdot|\boldsymbol{\theta}_k), \boldsymbol{\theta}_k = (\theta_1, \theta_2, \ldots, \theta_k, 0, \ldots, 0)\} \quad 1 \leq k \leq K$$

avec $\boldsymbol{\theta}_k$ paramètres du sous-espace $E_k \subset R^k$. Chaque modèle $\mathcal{M}_k$ a k paramètres libres et est une restriction du modèle $\mathcal{M}_{k+1}$. $\mathcal{M}_K$ est le modèle le plus général.

Entropie et information de kullback-Leibler

L'information de kullback-Leibler où l'entropie de Boltzmann entre les deux densités $f(\cdot|\boldsymbol{\theta}^*)$ et $f(\cdot|\boldsymbol{\theta})$ est définie par :

$$\mathrm{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) = \int f(u|\boldsymbol{\theta}^*)\log f(u|\boldsymbol{\theta}^*)du - \int f(u|\boldsymbol{\theta}^*)\log f(u|\boldsymbol{\theta})du$$

Cette quantité représente la perte d'information lorsque l'on choisit le paramètre $\boldsymbol{\theta}$ pour $f(\cdot|\boldsymbol{\theta})$ alors que le vrai paramètre est $\boldsymbol{\theta}^*$. Cette quantité est généralement appelée *cout entropique*. Notons quelques propriétés fondamentales de $\mathrm{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$ développées par Kullback(1959).

- $\mathrm{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) \geq 0$
- $\mathrm{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) = 0, \quad ssi \quad f(\cdot|\boldsymbol{\theta}) = f(\cdot|\boldsymbol{\theta}^*) \quad p.s.$
- lorsque $X_1,\ldots,X_n$ sont i.i.d. alors l'information de Kullback-Leibler pour l'échantillon $\mathcal{X}$ est $\mathrm{I}_n(\boldsymbol{\theta}^*;\boldsymbol{\theta}) = n\mathrm{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$ (propriété additive).

Le premier terme dans $\mathrm{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$ étant invariant selon les modèles considérés, minimiser $\mathrm{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$ revient à maximiser :

$$\mathrm{KL}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) = \int f(u|\boldsymbol{\theta}^*)\log f(u|\boldsymbol{\theta})du.$$

Principe d'Akaike

Akaike [1] propose de retenir le modèle $\mathcal{M}_k$ qui maximise l'espérance de la log vraisemblance :

$$\mathrm{E}\left[\log f(X|\boldsymbol{\theta}_k)\right] = \mathrm{E}\left[\int f(u|\boldsymbol{\theta}^*)\log f(u|\widehat{\boldsymbol{\theta}}_k)du\right]$$

où $\widehat{\theta}_k$ est l'estimateur du maximum de vraisemblance du paramètre de la densité f sous les contraintes de ce modèle et X est une variable aléatoire de densité $f(\cdot|\boldsymbol{\theta}^*)$ et est indépendante de $\widehat{\boldsymbol{\theta}}_k$. Ce qui revient à minimiser $\mathrm{E}\left[\mathrm{I}(\boldsymbol{\theta}^*;\widehat{\boldsymbol{\theta}})\right]$.
Cette quantité n'est pas directement calculable et nécessite d'être estimée.

Estimation de $\mathrm{E}\left[\mathrm{I}(\boldsymbol{\theta}^*;\widehat{\boldsymbol{\theta}})\right]$

Un développement de Taylor à l'ordre 2 de $\mathrm{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta})$ au voisinage de $\boldsymbol{\theta}^*$ donne :

$$\mathrm{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) = \underbrace{\mathrm{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta}^*)}_{0} + \underbrace{\left[\frac{\partial}{\partial\boldsymbol{\theta}}\mathrm{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta})\right]'_{|\boldsymbol{\theta}=\boldsymbol{\theta}^*}}_{0}(\boldsymbol{\theta}-\boldsymbol{\theta}^*) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)'\,\mathbf{J}\,(\boldsymbol{\theta}-\boldsymbol{\theta}^*)$$

avec $\mathbf{J}$ la matrice de Fischer $(K\times K)$ pour une seule donnée

$$\mathbf{J} = \mathrm{E}\left\{\left[\frac{\partial}{\partial\theta}\log f(X|\boldsymbol{\theta})\right]'\left[\frac{\partial}{\partial\theta}\log f(X|\boldsymbol{\theta})\right]_{|\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right\}.$$

Finalement on obtient

$$\mathrm{I}(\boldsymbol{\theta}^*;\boldsymbol{\theta}) \simeq \frac{1}{2}\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2_{\mathbf{J}}$$

avec $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^2_{\mathbf{J}} = \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}^*)'\,\mathbf{J}\,(\boldsymbol{\theta}-\boldsymbol{\theta}^*)$.
Soit $\boldsymbol{\theta}^*_k$ la projection de $\boldsymbol{\theta}^*$ sur le sous-espace E_k et $\widehat{\boldsymbol{\theta}}_k$ l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}^*_k$ dans E_k. Alors on a :

$$2\mathrm{I}(\boldsymbol{\theta}^*;\widehat{\boldsymbol{\theta}}_k) \simeq \|\boldsymbol{\theta}^*-\widehat{\boldsymbol{\theta}}_k\|^2_{\mathbf{J}}$$

Et en utilisant le théorème de pythagore on obtient :

$$2\mathrm{I}(\boldsymbol{\theta}^*;\widehat{\boldsymbol{\theta}}_k) \simeq \|\boldsymbol{\theta}^*-\boldsymbol{\theta}^*_k\|^2_{\mathbf{J}} + \|\boldsymbol{\theta}^*_k-\widehat{\boldsymbol{\theta}}_k\|^2_{\mathbf{J}}$$

En prenant n fois l'espérance on a :

$$\begin{aligned} 2n\mathrm{E}\left[\mathrm{I}(\boldsymbol{\theta}^*;\widehat{\boldsymbol{\theta}}_k)\right] &\simeq \mathrm{E}\left[n\|\boldsymbol{\theta}^*-\boldsymbol{\theta}^*_k\|^2_{\mathbf{J}} + n\|\boldsymbol{\theta}^*_k-\widehat{\boldsymbol{\theta}}_k\|^2_{\mathbf{J}}\right] \\ &= n\|\boldsymbol{\theta}^*-\boldsymbol{\theta}^*_k\|^2_{\mathbf{J}} + \mathrm{E}\left[n\|\boldsymbol{\theta}^*_k-\widehat{\boldsymbol{\theta}}_k\|^2_{\mathbf{J}}\right] \\ &= \delta + \mathrm{E}\left[\|n^{1/2}(\boldsymbol{\theta}^*_k-\widehat{\boldsymbol{\theta}}_k)\|^2_{\mathbf{J}}\right] \end{aligned}$$

avec $\delta = n\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_k^*\|_{\mathbf{J}}^2$. Ce terme δ représente le biais introduit par la projection de $\boldsymbol{\theta}^*$ sur E_k et le second terme est une mesure de la variance de l'erreur aléatoire $(\boldsymbol{\theta}_k^* - \widehat{\boldsymbol{\theta}}_k)$. Comme l'estimateur du maximum de vraisemblance $\widehat{\boldsymbol{\theta}}_k$ suit asymptotiquement une loi normale de moyenne $\widehat{\boldsymbol{\theta}}_k$ et de matrice de variance l'inverse de la matrice d'information de Fischer (pour l'echantillon complet) soit $(n\mathbf{J})^{-\mathbf{1}}$ on obtient :

$$\|n^{1/2}(\boldsymbol{\theta}_k^* - \widehat{\boldsymbol{\theta}}_k)\|_{\mathbf{J}}^2 \sim \chi_k^2.$$

Puisque $\mathrm{E}[\chi_k^2] = k$, on a

$$2n\mathrm{E}\left[\mathrm{I}(\boldsymbol{\theta}^*; \widehat{\boldsymbol{\theta}}_k)\right] \simeq \delta + k. \tag{A.1}$$

Il est encore impossible d'estimer δ qui est inconnue mais déterministe. Akaike [1] utilise le rapport du maximum de vraissemblance suivant :

$$\begin{aligned} -2\log R(\mathcal{X}) &= -2\log \frac{f(x_1,\ldots,x_n|\widehat{\boldsymbol{\theta}}_k)}{f(x_1,\ldots,x_n|\widehat{\boldsymbol{\theta}}_K)} \\ &= -2\sum_{i=1}^{n} \log \frac{f(x_i|\widehat{\boldsymbol{\theta}}_k)}{f(x_i|\widehat{\boldsymbol{\theta}}_K)} = -2(\mathrm{L}(\widehat{\boldsymbol{\theta}}_k) - \mathrm{L}(\widehat{\boldsymbol{\theta}}_K)) \end{aligned}$$

où $\mathrm{L}(\widehat{\boldsymbol{\theta}})$ représente la fonction de log-vraisemblance pour l'échantillon $\mathcal{X} = (x_1,\ldots,x_n)$. Wald [3] a montré que $-2\log R(\mathcal{X})$ a pour distribution asymptotique χ_{K-k}^2, un chi-deux décentré avec $(K-k)$ degré de liberté et de paramètre de décentrage $\delta = n\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_k^*\|_{\mathbf{J}}^2$

$$-2\log R(\mathcal{X}) \sim \chi_{K-k}^2(\delta)$$

Comme $\mathrm{E}[\chi_{K-k}^2(\delta)] = \delta + K - k$, on obtient

$$\begin{aligned} -2\log R(\mathcal{X}) &\simeq \mathrm{E}[2\log R(\mathcal{X})] \\ &\simeq \mathrm{E}[\chi_{K-k}^2(\delta)] \\ &\simeq \delta + K - k \end{aligned}$$

Finalement $\delta = -2\log R(\mathcal{X}) + k - K$. En combinant avec A.1, on trouve que

$$\begin{aligned} 2n\mathrm{E}\left[\mathrm{I}(\boldsymbol{\theta}^*;\widehat{\boldsymbol{\theta}}_k)\right] &= -2\log R(\mathcal{X}) + 2k - K \\ &= -2(\mathrm{L}(\widehat{\boldsymbol{\theta}}_k) - \mathrm{L}(\widehat{\boldsymbol{\theta}}_K)) + 2k - K \end{aligned}$$

Dans une procédure de sélection, on supprime les termes invariants et on obtient ainsi le critère AIC :

$$\mathrm{AIC}(\mathcal{M}_k) = -2\mathrm{L}(\widehat{\boldsymbol{\theta}}_k) + 2k$$

Le modèle choisit est celui qui réalise un AIC minimum.

Bibliographie

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

[2] H. Bozdogan. Model selection and Akaike's information criterion (AIC) : the general theory and its analytical extensions. *Psychometrika*, pages 345–370, 1987.

[3] A. Wald. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54 :426–482, 1987.

Publications de l'auteurs

[1] Liquet B, Commenges D. *Correction of the P-value after multiple coding of an explanatory variable in logistic regression.* Stat. Med. 2001 ; 20

[2] Liquet B, Sakarovitch C, Commenges D. *Bootstrap choice of estimators in parametric and semi-parametric families : an extension of EIC.* Biometrics. 2002 (sous presse)

[3] Liquet B, Commenges D. *Estimating the expectation of the log-likelihood with incomplete data for choosing an estimator in semi-parametric families.* Biometrics. 2002 (soumis)

[4] Liquet B, Sutanto *The impact of changing the rules in badminton.* American Mathematical Monthly. 2002 (soumis)

Communications dans un congrès avec comité de lecture :

[5] Liquet B, Commenges D. *Correction de la P-value après codage multiple d'une variable explicative dans un modèle de régression logistique.* XXXIIe Journées de Statistique, 15-19 mai 2000, Fes.

[6] Liquet B, Commenges D. *Sélection de modèles de régression semi-paramétrique par un critère d'information.* XXXIIIe Journées de Statistique, 14-18 mai 2001, Nantes.

[7] Liquet B, Sutanto *The impact of changing the rules in badminton.* First International Conference in Bordeaux on Statitical Methods for Olympic Data, 5-7 September 2001, Bordeaux.

[8] Liquet B, Commenges D. *Sélection de modèles de régression paramétrique et non-paramétrique par un critère d'information.* XXXIVe Journées de Statistique, 13-17 mai 2002, Bruxelles.