Sélection de variables dans le modèle de Cox

Les trois travaux présentés dans cette partie ont le même cadre et le même but : on cherche à expliquer une durée X, éventuellement censurée, par un certain nombre de grandeurs appelées covariables. Le modèle le plus utilisé pour traiter ce type de données est le modèle de Cox, dans lequel le taux de risque conditionnel à la covariable Z du temps de survie X s'écrit sous la forme

$$\alpha_Z(t) = e^{f(Z)}\alpha_0(t),\ t > 0,$$

où f est la fonction de régression à estimer et α_0 est le taux de risque de base, considéré ici comme un paramètre de nuisance. Il s'agit alors d'estimer au mieux la fonction f qui lie la covariable Z à la réponse X ou de sélectionner les covariables les plus pertinentes (c'est-à-dire les composantes de $Z \in \mathbb{R}^p$), à partir d'un n-échantillon du triplet (T, δ, Z), où T est la durée censurée et δ l'indicateur de censure : $T = \min(X, U),\ \delta = \mathbb{1}_{\{X \leq U\}}$, où U est la variable aléatoire de censure.

Les trois travaux diffèrent par les critères utilisés pour atteindre le but et par la nature des résultats proposés.

1. Estimation de la fonction de régression dans le modèle de Cox non paramétrique
   Dans le cadre de ma thèse [T], en collaboration avec Gwénaëlle Castellan (Université de Lille), nous avons suivi la procédure de sélection de modèles développée par Birgé et Massart [BarBirMas] pour estimer la fonction f de manière non paramétrique. Nous avons proposé un estimateur de la fonction de régression et montré que son risque est de l'ordre du plus petit risque des estimateurs construits dans une collection de modèles.
   Ce travail a été présenté dans plusieurs séminaires [S-Let02b, S-Let01, S-Let00a, S-Let00b].
2. Le « Dantzig Selector » pour le modèle de Cox
   Dans le cadre de mes recherches au LJK, en collaboration avec Anestis Antoniadis (LJK) et Piotr Fryzlewicz (London School of Economics), nous avons adapté au cas du modèle de Cox une méthode de sélection de variables, appelée « Dantzig Selector » et développée par Candès et Tao dans le cadre de la régression gaussienne [CanTao]. Ici, la fonction de régression f est supposée linéaire, $f(Z) = \beta^T Z$, β et Z étant deux vecteurs de $\mathbb{R}^p$, avec p très supérieur à n. Notre estimateur est construit en minimisant la norme L_1 de β, sous la contrainte $\|U(\beta)\|_\infty \leq \gamma_{n,p}$, où U est le processus des scores (vecteur des dérivées de la log-vraisemblance partielle de Cox). Cette procédure force certains coefficients de β à s'annuler et permet donc de réaliser une vraie sélection de variables.
   Nos résultats sont à la fois théoriques (propriétés de notre estimateur), algorithmiques (proposition d'un algorithme efficace) et appliqués à des données de biopuces.
   Ce travail a fait l'objet d'un article paru dans le *Scandinavian Journal of Statistics* [A-AntFryLet] et d'un exposé dans un colloque spécialisé dans les données de génomique [C-AntFryLet].
3. Algorithme PLS pour le modèle de Cox
   Sophie Lambert-Lacroix et moi-même avons développé un algorithme de type PLS pour le modèle de Cox. PLS (Partial Least Squares) est une méthode de réduction de dimension (dans le cas $n << p$) qui consiste à rechercher une ou plusieurs combinaisons linéaires des covariables qui maximisent la covariance avec la réponse (initialement supposée continue). En suivant l'exemple de Fort et Lambert-Lacroix [ForLam] dans le cadre de la régression logistique, nous construisons une pseudo-réponse continue à partir de la log-vraisemblance partielle de Cox, linéaire par rapport aux covariables, nous lui appliquons l'algorithme PLS, en ajoutant de plus une pénalité de type Ridge pour assurer la convergence de l'algorithme. Nous avons appliqué notre procédure à plusieurs jeux de données et l'avons comparé à des méthodes concurrentes.
   Ce travail a fait l'objet d'un article soumis dans une revue internationale [F-LamLet11] et de deux exposés [E-LamLet11, G-LetLam10].

Références

Articles scientifiques dans des revues à comité de lecture

[A-AntFryLet] Antoniadis A., Fryzlewicz P. et Letué F. (2010) " The Dantzig selector in Cox's proportional hazards model." *Scand. J. Stat.,* **37**, (4), pp. 531–552.

[BarBirMas] Barron, A.R., Birgé, L., Massart, P. (1999). "Risk bounds for model selection via penalization." *Probab. Th. Rel. Fields* **113**, 301-413.

[CanTao] Candès, E. and Tao, T. (2007). "The Dantzig selector : Statistical estimation when p is much larger than n." *Annals of Statistics*, **35** 2313-2351.

[ForLam] Fort, G. and Lambert-Lacroix, S. (2005) "Classification using partial least squares with penalized logistic regression", *Bioinformatics*, **21**, 7, 1104–1111.

Conférences

[E-LamLet11] Lambert-Lacroix S. et Letué F. (2011) ≪ PLS et modèle de Cox avec application aux données d'expression de gènes ≫, In 43^{emes} journées de Statistique, Tunis, Tunisie,

[C-AntFryLet] Antoniadis A., Fryzlewicz P. et Letué F. (2009) " The Dantzig selector in Cox's proportional hazards model. " In Statistical Methods for Post-genomic Data workshop, SMPGD'09. Paris, France.

Thèse

[T] Letué F. (2000) ≪ Modèle de Cox : Estimation par sélection de modèle et modèle de chocs bivarié. ≫ Thèse n^o 6414 de l'Université de Paris XI Orsay, décembre 2000.

Rapports techniques

[F-LamLet11] Lambert-Lacroix S. et Letué F. (2011) "Partial Least Squares and Cox model with application to gene expression", Rapport technique HAL.

Séminaires

[G-LetLam10] Letué F. et Lambert-Lacroix, S. (2010) ≪ PLS dans le modèle de Cox et application aux données d'expression de gènes ≫, Rencontres statistiques lyonnaises.

[S-Let02b] Letué F. (2002) ≪ Estimation de la fonction de régression de Cox non paramétrique par sélection de modèle. ≫Séminaire commun LMC/LabSAD/INRIA de Statistique à Grenoble.

[S-Let01] Letué F. (2001) ≪ Estimation de la fonction de régression de Cox non paramétrique par sélection de modèle. ≫ Séminaire de Statistiques de l'Université Paul Sabatier Toulouse III

[S-Let00a] Letué F. (2000) ≪ Estimation de la fonction de régression de Cox non paramétrique par sélection de modèle. ≫ Séminaire de Probabilités et Statistique de l'Université de Provence Aix-Marseille I.

[S-Let00b] Letué F. (2000) ≪ Estimation de la fonction de régression de Cox non paramétrique par sélection de modèle. ≫ Séminaire de Statistiques mathématiques et Applications, Garchy, France.