Estimation d'une mesure non paramétrique de l'importance d'une variable continue

Pierre Neuvial[1], Antoine Chambaz[2] & Mark J. van der Laan[3]

[1] *Laboratoire Statistique et Génome,*
Université d'Évry Val d'Essonne, UMR CNRS 8071 – USC INRA
pierre.neuvial@genopole.cnrs.fr
[2] *MAP5, Université Paris Descartes et CNRS*
antoine.chambaz@mi.parisdescartes.fr
[3] *Division of Biostatistics and Department of Statistics,*
University of California at Berkeley, Berkeley, CA 94720
laan@stat.berkeley.edu

Preprint : Chambaz et al. [2011], http://hal.archives-ouvertes.fr/hal-00629899/en

Résumé. Nous proposons une mesure non paramétrique de l'importance d'une variable continue, et développons une procédure d'estimation semi-paramétrique par minimisation de perte ciblée pour son inférence. Nous étudions ses propriétés théoriques (convergence de l'algorithme d'estimation, consistance et normalité asymptotique de l'estimateur) et son implémentation.

L'utilisation de cette procédure est illustrée à l'aide d'un exemple d'application à des données génomiques en cancérologie, qui est à l'origine de son développement : l'évaluation, pour chaque gène d'un échantillon tumoral, de l'influence du nombre de copies d'ADN sur le niveau d'expression du gène,en prenant en compte la méthylation de l'ADN dans la région promotrice du gène.

Mots-clés. Estimation non paramétrique, importance de variables, estimation ciblée, propriétés asymptotiques, robustesse, applications en biologie et en médecine.

Abstract. We define a new, non parametric measure of importance for a continuous variable. We develop a dedicated semi-parametric estimation procedure based on Targeted Minimum Loss Estimation (TMLE) methodology. Our work covers its theoretical study (convergence of the iterative procedure which is at the core of the TMLE methodology ; consistency and asymptotic normality of the estimator) and its implementation.

The application of this procedure is illustrated by a genomic data analysis example in cancer studies, which was the original motivation for this work. It consists in quantifying the influence of the DNA copy number on the expression level of the gene for a given cancer sample, taking into account the DNA methylation in the promoter region of the gene.

Keywords. Non parametric estimation, variable importance, targeted minimum loss estimation, asymptotic properties, robustness, applications to biology and medical sciences.

1 Une mesure de l'importance d'une variable

Motivation. On considère un gène donné dans l'ensemble des cellules d'un échantillon tumoral. On cherche à quantifier le degré d'association entre le nombre de copies d'ADN du gène (noté X) et son niveau d'expression (noté Y) conditionnellement au degré de méthylation de l'ADN (noté W) dans sa région promotrice. Une telle mesure d'association peut permettre d'identifier des gènes impliqués dans la progression tumorale, et qui n'auraient pas pu être repérés *via* une étude menée indépendamment sur X et sur Y [Pollack et al., 2002].

Notre objectif est de proposer une mesure d'association tenant compte des spécificités des données de nombre de copies d'ADN dans les cancers : d'une part l'existence d'un niveau de référence ($x_0 = 2$ copies d'ADN dans une cellule normale), et d'autre part le fait que les nombres de copies d'ADN observés dans les cellules tumorales présentent typiquement un continuum de valeurs (notamment car un échantillon dit "tumoral" est généralement un mélange de cellules normales et de cellules tumorales, et que les cellules tumorales elles-mêmes peuvent être de plusieurs sous-types tumoraux aux altérations spécifiques).

Définition du paramètre d'intérêt. Plus généralement, on note $O = (W, X, Y)$ un triplet d'intérêt, où $W \in \mathcal{W}$ est un vecteur de covariables, et X et Y sont deux mesures réelles telles que la loi de X et continue, et a une masse positive en x_0 connu. Afin de faire le moins possibles d'hypothèses (nécessairement réductrices) sur la loi P_0 de O, on suppose simplement que P_0 est un élément de l'ensemble non paramétrique $\mathcal{M}$ de toutes les lois candidates P pour O telles que $P(X \neq x_0) > 0$ et $P(X \neq x_0|W) > 0$ P-presque sûrement. On définit le paramètre d'intérêt $\psi_0 = \Psi(P_0)$, où $\Psi : \mathcal{M} \to \mathbb{R}$ est la fonctionnelle caractérisée par

$$\Psi(P) = \underset{\beta \in \mathbb{R}}{\arg\min}\, E_P\left\{(E_P(Y|X,W) - E_P(Y|X = x_0, W) - \beta(X - x_0))^2\right\} \qquad (1)$$

pour tout $P \in \mathcal{M}$. Le paramètre $\Psi(P)$ est une mesure de l'*importance* de X relativement à Y en prenant W en compte. De plus, il s'agit bien d'une mesure *non paramétrique* car elle est définie indépendamment de tout modèle semi-paramétrique, celui-ci prenant la forme $Y = \beta(X - x_0) + \eta(W)$ pour un paramètre de nuisance η non spécifié.

Interprétation du paramètre en termes d'excès de risque. On peut réécrite la fonctionnelle Ψ sous la forme suivante :

Proposition 1 *Pour tout $P \in \mathcal{M}$,*

$$\Psi(P) = \frac{E_P\{X(\theta(P)(X,W) - \theta(P)(0,W))\}}{E_P\{X^2\}}, \qquad (2)$$

où $\theta(P)(X,W) = E_P(Y|X,W)$.

La fonctionnelle Ψ peut s'interpréter comme une généralisation de la notion d'excès de risque à une mesure X continue. En effet, dans le cas où X ne prend que deux valeurs $x_0 = 0$ et x_1, alors on a $\Psi(P) = E_P\{(\theta(P)(x_1, W) - \theta(P)(0, W))h(P)(W)\}$, avec $h(P)(W) = P(X = x_1|W)/E_P\{X^2\}$; autrement dit, $\Psi(P)$ est une version pondérée par h de l'excès de risque classique $E_P\{\theta(P)(x_1, W) - \theta(P)(0, W)\}$.

2 Estimation par minimisation de perte ciblée

Principe général. Puisque la fonctionnelle Ψ est connue, on peut associer à tout estimateur P_n de la loi P des observations un estimateur de substitution de $\psi_n = \Psi(P_n)$. Nous proposons de mettre en place un estimateur reposant sur la théorie de la minimisation de perte ciblée (TMLE, pour Targeted Minimal Loss Estimation) qui est décrite en détail dans van der Laan and Rose [2011]. Cette théorie propose un schéma générique d'estimation itérative, que l'on peut résumer comme suit. La première étape consiste à construire un estimateur de substitution initial $\psi_n^0 = \Psi(P_n^0)$ à partir d'un estimateur initial P_n^0 de la loi P des observations. L'estimateur ψ_n^0 est ensuite utilisé pour construire une mise à jour de l'estimateur P_n^0, notée P_n^1, qui induit elle-même une mise à jour de l'estimateur de substitution $\psi_n^1 = \Psi(P_n^1)$.

Le cœur de l'approche TMLE réside dans l'étape de mise à jour. Celle-ci repose sur une notion fondamentale en statistique semi-paramétrique, la notion de *fonction d'influence efficace*, dont une définition mathématique est donnée par van der Vaart [1998]. Comme son nom l'indique, celle-ci quantifie l'influence de la loi P sur l'estimation du paramètre d'intérêt, et indique donc dans quelle *direction* l'estimation de la loi de P doit être modifiée pour que l'estimateur de substitution associé s'améliore (au sens de l'erreur quadratique). Pour $k \geq 0$, la mise à jour P_n^{k+1} de l'estimateur courant P_n^k de la loi des observations est définie comme suit. On considère un modèle paramétrique de dimension 1, de la forme $\{P_n^k(\varepsilon) : |\varepsilon| < \eta_n^k\}$, contenant P_n^k et dont le score en $\varepsilon = 0$ est la fonction d'influence efficace en P_n^k. Ce modèle est appelé *fluctuation* de la loi P_n^k. On définit alors $P_n^{k+1} = P_n^k(\varepsilon_n^k)$, où ε_n^k est l'estimateur du maximum de vraisemblance du modèle ci-dessus.

Algorithme de mise en oeuvre. L'étude théorique de la fonction d'influence efficace de Ψ révèle que les "traits" de la loi P qui importent pour l'estimation de ψ_0 sont $\theta(P)(X, W) = E_P(Y|X, W)$, $\mu(P)(W) = E_P(X|W)$, $g(P)(0|W) = P(X = 0|W)$, et $\sigma^2(P) = E_P\{X^2\}$. Pour l'étape d'initialisation, on forme des estimateurs de premier pas $(\theta_n^0, \mu_n^0, g_n^0, \sigma_n^0)$ des traits (θ, μ, g, σ) par super-learning, une méthode générique d'agrégation d'estimateurs par validation croisée proposée par van der Laan et al. [2007]. Afin de pouvoir ensuite estimer ψ_0 grâce à l'équation (2), on met en place une procédure d'estimation de la loi marginale de (W, X) reposant sur des simulations.

Pour la $k+1$-ème étape de mise à jour (avec $k \geq 0$), on estime la fonction d'influence efficace à l'aide des traits θ_n^k, μ_n^k, g_n^k et σ_n^k. Cette estimation détermine un modèle

paramétrique de dimension 1 (la fluctuation), dont le pas ε_n^k est déduit par maximisation de la vraisemblance. Ce pas optimal détermine à son tour la mise à jour P_n^{k+1} de l'estimation de la loi P. On met alors à jour en conséquence l'estimation des traits : θ_n^{k+1}, μ_n^{k+1}, g_n^{k+1} et σ_n^{k+1}, et finalement celle du paramètre : ψ_n^{k+1} grâce à de nouvelles simulations de la loi marginale de (X, W). On définit un critère d'arrêt de la procédure itérative, qui repose sur la décroissance des gains en vraisemblance successifs, ainsi que sur les variations des estimateurs associés.

Implémentation. La méthode décrite ci-dessus a été implémentée en langage R dans le paquet TMLE.NPVI qui est en cours de finition. L'estimation par super-learning a été réalisé grâce au paquet SuperLearner [Polley and van der Laan, 2011].

3 Résultats théoriques

Convergence de la procédure itérative. On peut montrer que si la suite (paramétrée par k) des moyennes quadratiques de la fonction d'influence efficace prise en P_n^k est écartée de 0, alors la procédure itérative proposée converge, au sens où la suite des ε_n^k tend vers 0. Par ailleurs, si la suite des ε_n^k tend vers 0 assez rapidement (au sens où la série $\sum_k |\varepsilon_n^k|$ converge), alors la suite des estimateurs $\psi_n^k = \Psi(P_n^k)$ converge également.

Propriétés asymptotiques de l'estimateur proposé. On peut montrer que l'estimateur TMLE proposé est consistant dès que les estimateurs des traits θ, μ, σ et g sont convergents et que l'on a soit la consistance de l'estimateur de $\theta(0, \cdot)$, soit celle des estimateurs de μ et g. Sous des conditions supplémentaires sur les vitesses de convergence des traits, le TMLE satisfait un théorème de la limite centrale. Enfin, si en outre les estimateurs de tous les traits sont consistants, alors le TMLE est efficace, et on dispose d'un estimateur de sa variance asymptotique.

4 Application en cancérologie

Nous avons appliqué la procédure d'estimation décrite à la Section 2 au problème biologique décrit à la Section 1 de recherche d'associations entre nombres de copies d'ADN et niveau d'expression d'un gène, conditionnellement à son degré de méthylation dans des échantillons tumoraux. Nous avons exploité des données provenant du consortium américain The Cancer Genome Atlas (TCGA).

Simulations. Nous avons dans un premier temps effectué une étude de simulation afin de tester l'implémentation de la méthode et d'illustrer les caractéristiques de l'estimateur proposé. Afin de rendre nos simulations aussi réalistes que possible, nous avons construit

un mécanisme de simulation consistant à perturber des données génomiques réelles de cancers du cerveau [The Cancer Genome Atlas (TGCA) research Network, 2008], concernant le gène EGFR, qui est connu pour être impliqué dans la progression tumorale de ces cancers. Les simulations que nous avons réalisées illustrent la robustesse, la rapidité de convergence, et la normalité asymptotique de l'estimateur proposé.

Analyse de données de cancers de l'ovaire Nous avons appliqué la méthode d'estimation proposée à des données de cancer de l'ovaire [The Cancer Genome Atlas (TGCA) research Network, 2011]. Le jeu de données est constitué de près de 500 échantillons tumoraux dont on a mesuré le nombre de copies d'ADN X, le niveau d'expression Y, et le degré de méthylation W à l'aide d'expériences de puces à ADN. Nous nous sommes focalisés sur 130 gènes situés sur le chromosome 18. Notre étude permet notamment d'isoler un groupe de 5 gènes situés dans la région 18q11.2 pour lesquels notre mesure d'association est particulièrement significative. Ceci suggère que cette région du génome pourrait être impliquée dans le processus tumoral.

Références

A. Chambaz, P. Neuvial, and M. J. van der Laan. Estimation of a non-parametric variable importance measure of a continuous exposure. Preprint, October 2011. URL http://hal.archives-ouvertes.fr/hal-00629899/en.

J. R. Pollack, T. Sørlie, C. M. Perou, C. A. Rees, S. S. Jeffrey, P. E. Lonning, R. Tibshirani, D. Botstein, A.-L. Børresen-Dale, and P. O Brown. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci U S A*, 99(20) :12963–12968, Oct 2002.

E. Polley and M. J. van der Laan. *SuperLearner*, 2011. URL http://CRAN.R-project.org/package=SuperLearner. R package version 2.0-4.

The Cancer Genome Atlas (TGCA) research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455 :1061–1068, 2008.

The Cancer Genome Atlas (TGCA) research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353) :609–615, 2011.

M. J. van der Laan and S. Rose. *Targeted Learning : Causal Inference for Observational and Experimental Data.* Springer Verlag, 2011.

M. J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Stat. Appl. Genet. Mol. Biol.*, 6 :Article 25, 2007.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.