Section 7

Étude de cas 2 : estimation ciblée de l'effet d'un biomarqueur

Lectures : Tuglus et van der Laan (2011), ch. 22 de *Targeted Learning* ; van der Laan et Rose (2011), ch. 4 et 5 ; notes de la section 6 (fonction d'influence)

Julien Miron – notes rédigées à partir du matériel de Lajmi Lakhal-Chaieb et Alexandre Bureau

Objectifs de la séance

1. Formuler la découverte de biomarqueurs comme un problème d'*estimation d'un paramètre cible* plutôt que comme un problème de sélection de modèle.
2. Définir l'effet marginal ajusté d'un biomarqueur continu A en présence de nombreuses covariables W (les autres gènes), dans un modèle semi-paramétrique partiellement linéaire, et en donner une interprétation non paramétrique.
3. Comprendre chaque étape de l'estimateur ciblé par maximum de vraisemblance (TMLE) de cet effet : estimateur initial, mécanisme de « traitement », covariable astucieuse, fluctuation, substitution.
4. Faire l'inférence à partir de la courbe d'influence, et comprendre les propriétés de double robustesse et d'efficacité locale.
5. Préparer l'étude de cas 2 (document de consignes de l'étude de cas 2), en particulier la justification des moindres carrés pour une réponse dichotomique et l'implantation avec des splines adaptatives (`earth`).

1 Mise en contexte : la découverte de biomarqueurs

1.1 Le problème

Une étude d'expression génique mesure, chez n sujets, l'expression de plusieurs milliers de gènes et une issue clinique Y (type de cancer, réponse au traitement, survie). On veut dresser une liste ordonnée de gènes « importants » pour Y, qui serviront de biomarqueurs (diagnostic, pronostic) ou de pistes biologiques. L'exemple du chapitre est le jeu de données de Golub et coll. (1999) classique : 38 échantillons de moelle osseuse de patients atteints de leucémie, 27 leucémies lymphoblastiques aiguës (ALL) et 11 leucémies myéloïdes aiguës (AML), avec l'expression de 3051 gènes (version du module `multtest`). La réponse est $Y = \mathbf{1}\{\text{ALL}\}$ (ou AML, selon le codage retenu).

Deux approches usuelles posent problème.

— **Les analyses marginales** (un test t par gène, puis contrôle de la FDR, section 11). Elles mesurent l'association brute entre chaque gène et Y. Or les gènes sont fortement corrélés (voies biologiques communes) : un gène peut être associé à Y uniquement parce qu'il est corrélé à un autre gène qui, lui, est associé à Y. Les autres gènes jouent le rôle de *facteurs de confusion*.

— **Les analyses multivariées par sélection de modèle** (régression pénalisée, sélection pas à pas, forêts aléatoires). Le coefficient d'un gène dépend alors de l'ensemble des autres gènes retenus et de la forme fonctionnelle imposée ; il n'a pas d'interprétation stable d'un modèle à l'autre. L'inférence après sélection est en général invalide, et les mesures d'« importance » des forêts aléatoires n'ont pas de loi d'échantillonnage connue.

L'idée de l'apprentissage ciblé (van der Laan et Rose, 2011) est de renverser l'ordre : on *définit d'abord* le paramètre qui répond à la question scientifique, comme une fonctionnelle de la loi des données qui garde le même sens quel que soit l'algorithme d'estimation ; on *estime ensuite* les quantités nuisibles par apprentissage machine, sans hypothèse paramétrique ; enfin on *cible* l'estimateur sur le paramètre d'intérêt pour retrouver une loi asymptotique normale et une erreur type valide.

1.2 La feuille de route de l'apprentissage ciblé

1. **Données** : $O_1, \ldots, O_n$ i.i.d. de loi P_0, avec $O = (W, A, Y)$.
2. **Modèle statistique** $\mathcal{M}$: ensemble des lois possibles pour P_0 ; aussi grand que les connaissances le permettent (non paramétrique ou semi-paramétrique).
3. **Paramètre cible** : une application $\Psi : \mathcal{M} \to \mathbb{R}^d$; la quantité d'intérêt est $\psi_0 = \Psi(P_0)$.
4. **Estimateur** : un estimateur de substitution $\Psi(P_n^*)$, où P_n^* est une estimation de (la partie pertinente de) P_0 ajustée pour bien estimer ψ_0.
5. **Inférence** : à partir de la courbe d'influence de l'estimateur.

Pour un gène donné, on note A son expression (le biomarqueur candidat), W le vecteur d'expression des autres gènes (ou d'un sous-ensemble choisi), et Y l'issue. On répète ensuite l'analyse pour chaque gène candidat, en jouant tour à tour le rôle de A.

2 Le paramètre cible

2.1 Notation

On note

$$\bar{Q}_0(A,W) = \mathrm{E}_0[Y \mid A, W], \qquad g_0(W) = \mathrm{E}_0[A \mid W], \qquad \sigma_0^2(A,W) = \mathrm{Var}_0(Y \mid A, W).$$

Par analogie avec l'inférence causale, g_0 est appelé *mécanisme de traitement*, même si un gène n'est pas une intervention. Quand Y est dichotomique, $\bar{Q}_0(A,W) = \mathbb{P}_0(Y = 1 \mid A, W)$ et $\sigma_0^2 = \bar{Q}_0(1 - \bar{Q}_0)$.

2.2 Biomarqueur dichotomique : la différence de risque ajustée

Si $A \in \{0,1\}$, le paramètre étudié au chapitre 4 de van der Laan et Rose (2011) est

$$\psi_0 = \mathrm{E}_W[\bar{Q}_0(1,W) - \bar{Q}_0(0,W)].$$

On compare, *à W fixé*, la moyenne de Y entre les deux niveaux de A, puis on fait la moyenne sur la loi marginale de W dans la population. Sous les hypothèses causales usuelles (absence de confusion non mesurée, positivité $0 < \mathbb{P}(A = 1 \mid W) < 1$), ψ_0 est un effet causal moyen. En découverte de biomarqueurs, on n'invoque pas d'interprétation causale : ψ_0 est une mesure d'*importance de variable* (van der Laan, 2006), c'est-à-dire l'association de A avec Y nette de celle qui transite par W.

2.3 Biomarqueur continu : l'effet linéaire marginal

L'expression génique est continue. La généralisation naturelle est la courbe

$$\psi_0(a) = \mathrm{E}_W[\bar{Q}_0(a,W) - \bar{Q}_0(0,W)], \qquad a \in \mathbb{R},$$

qui compare le niveau a au niveau de référence 0 (en pratique, les expressions sont centrées, de sorte que 0 est un niveau typique). Une courbe entière est difficile à estimer et à rapporter pour des milliers de gènes ; Tuglus et van der Laan la résument par un seul nombre en postulant un modèle semi-paramétrique.

Définition 2.1 (Modèle semi-paramétrique de l'effet d'un biomarqueur)**.** Le modèle $\mathcal{M}$ est l'ensemble des lois de $O = (W,A,Y)$ telles que

$$\bar{Q}(a,W) - \bar{Q}(0,W) = \beta a \quad \text{pour tout } a \text{ et presque tout } W, \tag{1}$$

la loi de W, la loi conditionnelle de A sachant W et le reste de la loi de Y sachant (A,W) n'étant pas restreints. Le paramètre cible est $\Psi(P) = \beta$; alors $\psi(a) = \beta a$.

En posant $r(W) = \bar{Q}(0,W)$, la contrainte (1) équivaut au *modèle partiellement linéaire*

$$\mathrm{E}[Y \mid A,W] = \beta A + r(W), \qquad r \text{ quelconque.} \tag{2}$$

La seule hypothèse est que l'effet de A est linéaire et ne varie pas avec W ; la dépendance en W, qui contient des centaines de gènes, est laissée libre. On peut généraliser à $\bar{Q}(a,W) - \bar{Q}(0,W) = a\,\beta^\top V$, où $V \subset W$ contient des modificateurs d'effet ; le chapitre mentionne cette extension, que nous laissons de côté.

En prenant l'espérance conditionnelle de (2) sachant W, on obtient $\mathrm{E}[Y \mid W] = \beta g(W) + r(W)$, d'où les deux identités utiles

$$r(W) = \mathrm{E}[Y \mid W] - \beta\, g(W), \qquad \mathrm{E}[Y \mid A,W] - \mathrm{E}[Y \mid W] = \beta\,\{A - g(W)\}. \tag{3}$$

La seconde identité est le point de départ de Robinson (1988) : après avoir retiré de Y et de A leur partie expliquée par W, il reste une régression linéaire simple sans ordonnée à l'origine,

$$\beta = \frac{\mathrm{E}[\{A - g(W)\}\{Y - \mathrm{E}(Y \mid W)\}]}{\mathrm{E}[\{A - g(W)\}^2]} = \frac{\mathrm{E}[\mathrm{Cov}(A,Y \mid W)]}{\mathrm{E}[\mathrm{Var}(A \mid W)]}. \tag{4}$$

2.4 Que vise-t-on si le modèle est faux ?

Le membre de droite de (4) est bien défini pour *toute* loi telle que $\mathrm{E}[\mathrm{Var}(A \mid W)] > 0$, que le modèle (1) soit vrai ou non. On peut donc prendre

$$\beta^*(P) = \frac{\mathrm{E}_P[\mathrm{Cov}_P(A, Y \mid W)]}{\mathrm{E}_P[\mathrm{Var}_P(A \mid W)]}$$

comme définition non paramétrique du paramètre : il coïncide avec β dans $\mathcal{M}$. Si, par exemple, $\bar{Q}(A,W) = \beta(W)A + r(W)$ (effet linéaire mais modifié par W), alors $\mathrm{Cov}(A,Y \mid W) = \beta(W)\,\mathrm{Var}(A \mid W)$ et

$$\beta^* = \mathrm{E}\big[\omega(W)\,\beta(W)\big], \qquad \omega(W) = \frac{\mathrm{Var}(A \mid W)}{\mathrm{E}[\mathrm{Var}(A \mid W)]},$$

une moyenne des pentes conditionnelles pondérée par la variabilité résiduelle du biomarqueur. C'est dans cet esprit que van der Laan (2006) définit les paramètres d'importance de variable par projection d'un paramètre non paramétrique sur un modèle de travail : le modèle linéaire sert à *résumer* l'effet, pas à le contraindre. Nous verrons (encadré « Niveau doctorat ») que la courbe d'influence utilisée pour l'inférence est la même dans les deux interprétations.

2.5 Identifiabilité et choix de W

Les formules (4) montrent que β n'est identifiable que si $\mathrm{E}[\mathrm{Var}(A \mid W)] > 0$: il faut que le biomarqueur varie encore une fois les autres gènes fixés. C'est l'analogue continu de l'hypothèse de positivité. En pratique, si W contient un gène presque colinéaire à A, $\mathrm{Var}(A \mid W)$ est très petite, l'information sur β s'effondre et l'estimateur devient instable. Tuglus et van der Laan limitent donc l'ajustement aux gènes dont la corrélation avec A est inférieure à un seuil (0,5 dans l'étude de cas). Ce choix modifie le paramètre (on ajuste pour moins de facteurs de confusion), mais il est nécessaire pour que le problème soit bien posé avec $n = 38$.

À retenir

- Le paramètre cible est défini *avant* le choix d'un algorithme : $\psi(a) = \mathrm{E}_W[\bar{Q}(a,W) - \bar{Q}(0,W)]$, résumé par β dans le modèle partiellement linéaire $\mathrm{E}[Y \mid A,W] = \beta A + r(W)$.
- $\beta = \mathrm{E}[\mathrm{Cov}(A,Y \mid W)] / \mathrm{E}[\mathrm{Var}(A \mid W)]$: c'est aussi un paramètre non paramétrique bien défini quand l'effet n'est pas linéaire.
- Il faut que A varie à W fixé (positivité) ; on exclut de W les gènes trop corrélés à A.

3 Rappel : courbe d'influence et équation estimante efficace

La section 6 a introduit les estimateurs asymptotiquement linéaires et la fonction d'influence efficace. Rappelons l'essentiel.

Définition 3.1. Un estimateur $\hat{\psi}_n$ de ψ_0 est *asymptotiquement linéaire* de courbe d'influence D si

$$\hat{\psi}_n - \psi_0 = \frac{1}{n}\sum_{i=1}^{n} D(O_i) + o_P(n^{-1/2}), \qquad \mathrm{E}_0[D(O)] = 0,\ \mathrm{E}_0[D(O)^2] < \infty.$$

Par le théorème central limite, $\sqrt{n}(\hat{\psi}_n - \psi_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{E}_0[D^2])$.

Dans un modèle $\mathcal{M}$ donné, la courbe d'influence de plus petite variance parmi celles des estimateurs réguliers est la *courbe d'influence efficace* (ou gradient canonique) D^* ; $\mathrm{E}_0[D^{*2}]$ est la borne d'efficacité semi-paramétrique (Tsiatis, 2006, ch. 4 ; van der Laan et Rose, 2011, annexe A1).

Pour le modèle partiellement linéaire (2), avec variance conditionnelle constante $\sigma^2(A,W) \equiv \sigma^2$, on montre (encadré « Niveau doctorat » de la section 7) que

$$D^*(O) = \frac{\{A - g_0(W)\}\{Y - \beta_0 A - r_0(W)\}}{\mathrm{E}_0\big[\{A - g_0(W)\}^2\big]}. \tag{5}$$

Le numérateur est le produit de deux résidus : celui du biomarqueur après prédiction par W, et celui de la réponse dans le modèle. Le dénominateur normalise. Lorsque la variance conditionnelle n'est pas constante (c'est le cas pour Y dichotomique), la fonction d'influence efficace fait intervenir des poids $1/\sigma^2(A,W)$; (5) reste alors une courbe d'influence valide, efficace seulement sous l'hypothèse de travail d'homoscédasticité. On parle d'efficacité *locale*.

3.1 L'équation estimante et sa double robustesse

La forme (5) suggère l'équation estimante

$$\sum_{i=1}^{n} \{A_i - g(W_i)\}\{Y_i - \beta A_i - r(W_i)\} = 0, \tag{6}$$

qui dépend de deux fonctions nuisibles, g et r.

Proposition 3.2 (Double robustesse)**.** *Soit* $\Phi(\beta, r, g) = \mathrm{E}_0\big[\{A - g(W)\}\{Y - \beta A - r(W)\}\big]$. *Sous le modèle* (2)*, pour toutes fonctions* r *et* g*,*

$$\Phi(\beta_0, r, g) = \mathrm{E}_0\big[\{g_0(W) - g(W)\}\{r_0(W) - r(W)\}\big].$$

En particulier, $\Phi(\beta_0, r, g) = 0$ *dès que* $r = r_0$ ou $g = g_0$.

Démonstration. Conditionnellement à (A,W), $\mathrm{E}_0[Y - \beta_0 A - r(W) \mid A, W] = r_0(W) - r(W)$, d'où $\Phi(\beta_0, r, g) = \mathrm{E}_0[\{A - g(W)\}\{r_0(W) - r(W)\}]$. En conditionnant ensuite sur W, $\mathrm{E}_0[A - g(W) \mid W] = g_0(W) - g(W)$, ce qui donne le résultat. □

Cette identité est au cœur de tout ce qui suit : le biais dû à l'estimation des nuisances est un *produit* de deux erreurs. Si chacune converge plus vite que $n^{-1/4}$, ce qui est accessible pour des méthodes d'apprentissage machine raisonnables, leur produit est $o(n^{-1/2})$ et ne perturbe pas la loi asymptotique de $\hat{\beta}$.

3.2 Pourquoi ne pas simplement substituer ?

Un estimateur naïf consiste à ajuster $\bar{Q}_n$ par une méthode flexible, puis à en extraire la pente : par exemple β_n^0, coefficient de la régression sans ordonnée à l'origine de $\bar{Q}_n(A_i,W_i) - \bar{Q}_n(0,W_i)$ sur A_i. Cet estimateur de substitution hérite du compromis biais-variance de l'algorithme : la régularisation (choix des nœuds, élagage, pénalité) est réglée pour bien prédire Y, pas pour bien estimer β. Son biais est d'ordre $\mathrm{E}_0[D^*(\bar{Q}_n, g_0)]$, qui, en général, ne décroît pas plus vite que l'erreur de $\bar{Q}_n$ elle-même, donc pas assez vite pour une inférence à la vitesse $\sqrt{n}$. Il n'a pas non plus de courbe d'influence connue. Le TMLE corrige précisément ce terme en modifiant $\bar{Q}_n$ de façon que l'équation (6) soit résolue.

4 L'estimateur TMLE de l'effet linéaire

4.1 Principe général

Le TMLE part d'un estimateur initial $\bar{Q}_n^0$ et construit un *sous-modèle de fluctuation* $\{\bar{Q}_n^0(\varepsilon) : \varepsilon \in \mathbb{R}\}$ qui passe par $\bar{Q}_n^0$ en $\varepsilon = 0$, reste dans le modèle $\mathcal{M}$, et dont le score (pour la fonction de perte choisie) en $\varepsilon = 0$ engendre la composante de D^* qui concerne $\bar{Q}$. On estime ε en minimisant le risque empirique, on met à jour $\bar{Q}_n^1 = \bar{Q}_n^0(\varepsilon_n)$, on itère au besoin, puis on substitue : $\hat{\psi} = \Psi(\bar{Q}_n^*)$. La mise à jour résout, par construction, l'équation $\frac{1}{n}\sum_i D^*(O_i) = 0$ pour les nuisances estimées.

Pour l'effet linéaire, la perte est la perte quadratique $\mathcal{L}(\bar{Q})(O) = \{Y - \bar{Q}(A,W)\}^2$, la fluctuation est linéaire et la covariable astucieuse est

$$H(A,W) = A - g(W), \tag{7}$$

c'est-à-dire la dérivée $\partial m(A,W \mid \beta)/\partial\beta = A$ du modèle $m(A,W \mid \beta) = \beta A$, moins sa projection sur les fonctions de W.

4.2 L'algorithme

Algorithme TMLE de l'effet linéaire β (notre présentation)

Étape 0. Choix de W. Retenir comme covariables d'ajustement les autres gènes dont la corrélation avec A est inférieure au seuil (positivité).

Étape 1. Estimateur initial. Ajuster $\bar{Q}_n^0(A,W)$, estimation de $\mathrm{E}[Y \mid A,W]$, par une méthode non paramétrique (splines polynomiales adaptatives `polspline`, `earth`, ou un super apprenant). Si Y est dichotomique, utiliser une version adaptée aux proportions (lien logit). En déduire

$$\beta_n^0 = \frac{\sum_i A_i\{\bar{Q}_n^0(A_i,W_i) - \bar{Q}_n^0(0,W_i)\}}{\sum_i A_i^2},$$

projection de $\bar{Q}_n^0(A,W) - \bar{Q}_n^0(0,W)$ sur $\{\beta A\}$, puis $r_n^0(W)$, estimation de $\mathrm{E}[Y - \beta_n^0 A \mid W]$ (variable réelle). L'estimateur initial *dans le modèle* est $\bar{Q}_n^0(A,W) = \beta_n^0 A + r_n^0(W)$.

Étape 2. Mécanisme de traitement. Estimer $g_n(W)$, prédiction de A par W, par une méthode non paramétrique (variable réelle).

Étape 3. Covariable astucieuse. Calculer $H_n(A_i,W_i) = A_i - g_n(W_i)$.

Étape 4. Fluctuation. Régresser Y sur H_n par moindres carrés, sans ordonnée à l'origine, avec $\bar{Q}_n^0(A,W)$ en décalage (*offset*) :

$$\begin{aligned}\varepsilon_n &= \arg\min_{\varepsilon} \sum_i \{Y_i - \bar{Q}_n^0(A_i,W_i) - \varepsilon H_n(A_i,W_i)\}^2 \\ &= \frac{\sum_i H_n(A_i,W_i)\{Y_i - \bar{Q}_n^0(A_i,W_i)\}}{\sum_i H_n(A_i,W_i)^2}.\end{aligned}$$

Étape 5. Mise à jour et substitution. $\bar{Q}_n^1 = \bar{Q}_n^0 + \varepsilon_n H_n$, c'est-à-dire

$$\beta_n^1 = \beta_n^0 + \varepsilon_n, \qquad r_n^1(W) = r_n^0(W) - \varepsilon_n\, g_n(W).$$

Le TMLE est $\hat{\beta} = \beta_n^1 = \Psi(\bar{Q}_n^1)$.

Étape 6. Inférence. Courbe d'influence estimée $\mathrm{IC}_n(O_i) = H_n(A_i,W_i)\{Y_i - \hat{\beta}A_i - r_n^1(W_i)\}/(\frac{1}{n}\sum_j H_n(A_j,W_j)^2)$; erreur type $\widehat{\mathrm{ET}} = \mathrm{sd}(\mathrm{IC}_n)/\sqrt{n}$; test de Wald $Z = \hat{\beta}/\widehat{\mathrm{ET}}$ comparé à la loi $\mathcal{N}(0,1)$.

Remarque 4.1. Le chapitre présente ces étapes dans deux encadrés gris (sous-section *Implementation*, p. 370–371) et note la fonction mise à jour $r_n^1(W)$; notre notation peut différer légèrement de celle du livre (voir le document de consignes). L'estimation initiale de r peut aussi se faire directement par $\bar{Q}_n^0(0,W)$; passer par une régression de $Y - \beta_n^0 A$ sur W évite d'extrapoler $\bar{Q}_n^0$ au point $A = 0$ et garantit que l'estimateur initial appartient au modèle (2).

4.3 Propriétés de la mise à jour

La fluctuation reste dans le modèle. Pour tout ε,

$$\bar{Q}_n^0(A,W) + \varepsilon H_n(A,W) = (\beta_n^0 + \varepsilon)A + \{r_n^0(W) - \varepsilon g_n(W)\},$$

qui est encore de la forme $\beta A + r(W)$. Le paramètre de fluctuation ε ne fait que déplacer la pente, et la fonction r absorbe la partie $-\varepsilon g_n(W)$: c'est pourquoi l'estimateur de substitution après mise à jour est simplement $\beta_n^0 + \varepsilon_n$.

Le score de la fluctuation est la courbe d'influence efficace. La dérivée de la perte quadratique en $\varepsilon = 0$ vaut

$$-\frac{1}{2}\frac{\partial}{\partial\varepsilon}\{Y - \bar{Q}_n^0 - \varepsilon H_n\}^2\Big|_{\varepsilon=0} = H_n(A,W)\{Y - \bar{Q}_n^0(A,W)\},$$

qui est, à la constante de normalisation près, D^* évaluée en $(\bar{Q}_n^0, g_n)$.

Une seule itération suffit. La covariable H_n ne dépend que de g_n, qui n'est pas modifié par la mise à jour. Après une étape, la condition du premier ordre des moindres carrés donne exactement

$$\sum_{i=1}^n H_n(A_i,W_i)\{Y_i - \bar{Q}_n^1(A_i,W_i)\} = \sum_{i=1}^n \{A_i - g_n(W_i)\}\{Y_i - \hat{\beta}A_i - r_n^1(W_i)\} = 0, \tag{8}$$

c'est-à-dire l'équation estimante (6) avec les nuisances estimées. Une seconde itération donnerait $\varepsilon = 0$. (Avec une fluctuation logistique, comme au chapitre 5 pour A dichotomique, la covariable dépend aussi de g_n seulement, et une étape suffit également ; ce n'est pas le cas pour tous les paramètres.)

Lien avec l'estimateur de Robinson. Si l'on remplace r_n^1 par $\hat{m}(W) - \hat{\beta}g_n(W)$, où $\hat{m}$ estime $\mathrm{E}[Y \mid W]$, l'équation (8) donne $\hat{\beta} = \sum_i H_n\{Y_i - \hat{m}(W_i)\}/\sum_i H_n^2$, l'estimateur « résidu sur résidu » de Robinson (1988), aujourd'hui popularisé sous le nom de *double machine learning* (Chernozhukov et coll., 2018). Le TMLE a l'avantage d'être un estimateur de substitution ($\hat{\beta} = \Psi(\bar{Q}_n^1)$ pour un $\bar{Q}_n^1$ du modèle) ; pour ce paramètre linéaire, les deux approches sont asymptotiquement équivalentes.

4.4 Les moindres carrés pour une réponse dichotomique

Dans l'application aux données de Golub, Y est dichotomique et le modèle (2) est un modèle *linéaire* de probabilité ; la fluctuation est estimée par moindres carrés. Plusieurs arguments justifient ce choix (question 1 de l'étude de cas ; les éléments ci-dessous sont des pistes à développer).

1. **La perte quadratique est valide pour une espérance conditionnelle.** Pour toute loi de Y, $\mathrm{E}[\{Y - f(A,W)\}^2] = \mathrm{E}[\{Y - \bar{Q}_0\}^2] + \mathrm{E}[\{\bar{Q}_0 - f\}^2]$, minimisée par $f = \bar{Q}_0$. Or, pour $Y \in \{0,1\}$, $\bar{Q}_0 = \mathbb{P}(Y = 1 \mid A,W)$: les moindres carrés visent bien la probabilité conditionnelle.
2. **L'équation estimante est sans biais quelle que soit la variance.** L'équation (6) n'utilise que $\mathrm{E}[Y \mid A,W]$; l'hétéroscédasticité $\sigma^2 = \bar{Q}(1 - \bar{Q})$ n'introduit pas de biais. Elle coûte seulement de l'efficacité : des moindres carrés pondérés par $1/\{\bar{Q}(1 - \bar{Q})\}$ seraient efficaces, au prix d'une instabilité quand $\bar{Q}$ s'approche de 0 ou 1. C'est exactement la situation des GEE avec matrice de travail indépendante, ou de la question de l'examen final H26 sur le modèle $\mathrm{E}[Y \mid X] = X^\top \beta$ pour Y dichotomique.
3. **La fluctuation linéaire respecte le modèle.** Une fluctuation logistique $\operatorname{logit} \bar{Q}_n^0(\varepsilon) = \operatorname{logit} \bar{Q}_n^0 + \varepsilon H_n$ produirait un effet de A qui n'est plus additif sur l'échelle des probabilités : on sortirait du modèle (1) et le paramètre ne serait plus simplement $\beta_n^0 + \varepsilon$.
4. **L'inférence ne repose pas sur la vraisemblance binomiale.** L'erreur type est calculée à partir de la variance empirique de la courbe d'influence, comme un estimateur sandwich ; elle est valide sans modéliser $\sigma^2(A,W)$.

Le prix à payer est que les prédictions $\bar{Q}_n^1$ peuvent sortir de $[0,1]$; ce n'est pas grave ici, puisque la cible est la pente β et non une prédiction individuelle.

À retenir

- TMLE = estimateur initial flexible + fluctuation dans la direction de la covariable astucieuse $H = A - g(W)$ + substitution.
- La fluctuation linéaire reste dans le modèle : $\hat{\beta} = \beta_n^0 + \varepsilon_n$ et $r_n^1 = r_n^0 - \varepsilon_n g_n$; une étape suffit.
- Après la mise à jour, l'équation estimante efficace est résolue exactement : c'est ce qui retire le biais de premier ordre de l'estimateur de substitution.
- Les moindres carrés sont légitimes pour une réponse dichotomique : ils visent $\mathbb{P}(Y = 1 \mid A,W)$ et l'hétéroscédasticité n'affecte que l'efficacité.

5 Propriétés asymptotiques et inférence

5.1 Normalité asymptotique

Théorème 5.1. *Supposons le modèle* (2) *vrai,* $\mathrm{E}_0[\mathrm{Var}(A \mid W)] > 0$*, et que les estimateurs* g_n *et* r_n^1 *satisfont*

(i) convergence *:* $\|g_n - g_\infty\|_2 \xrightarrow{\mathbb{P}} 0$ *et* $\|r_n^1 - r_\infty\|_2 \xrightarrow{\mathbb{P}} 0$ *pour des limites* (g_∞, r_∞) *dont au moins une est correcte* ($g_\infty = g_0$ *ou* $r_\infty = r_0$) *;*

(ii) condition de reste *:* $\|g_n - g_0\|_2 \, \|r_n^1 - r_0\|_2 = o_P(n^{-1/2})$ *;*

(iii) condition de complexité *: les fonctions estimées appartiennent à une classe de Donsker avec probabilité tendant vers* 1 *(ou l'on utilise l'ajustement croisé).*

Alors $\hat{\beta}$ *est asymptotiquement linéaire. Si* $g_\infty = g_0$ *et* $r_\infty = r_0$*, sa courbe d'influence est* D^* *en* (5) *et*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathrm{E}_0[D^{*2}]).$$

L'esquisse de la démonstration est donnée dans l'encadré « Niveau doctorat » (section 7). L'idée tient en une ligne : en écrivant (8) comme $P_n \phi(\hat{\beta}, r_n^1, g_n) = 0$ et en développant autour de β_0, on obtient

$$\hat{\beta} - \beta_0 = \underbrace{\frac{(P_n - P_0)\phi(\beta_0, r_0, g_0)}{\mathrm{E}_0[H_0^2]}}_{\text{terme linéaire}} + \underbrace{\frac{\Phi(\beta_0, r_n^1, g_n)}{\mathrm{E}_0[H_0^2]}}_{\text{produit d'erreurs}} + \underbrace{\frac{(P_n - P_0)\{\phi_n - \phi_0\}}{\mathrm{E}_0[H_0^2]}}_{\text{processus empirique}} + o_P(n^{-1/2}),$$

où le deuxième terme est contrôlé par la proposition 3.2, $P_n f = \frac{1}{n}\sum_i f(O_i)$, $P_0 f = \mathrm{E}_0[f(O)]$, $\phi_n = \phi(\beta_0, r_n^1, g_n)$, $\phi_0 = \phi(\beta_0, r_0, g_0)$ et $\phi(\beta,r,g)(O) = \{A - g(W)\}\{Y - \beta A - r(W)\}$.

5.2 Double robustesse et efficacité locale

— **Double robustesse.** Par la proposition 3.2, $\hat{\beta}$ est convergent si g_n converge vers g_0 *ou* si r_n^1 converge vers r_0. En découverte de biomarqueurs, c'est surtout g_0 (prédire un gène par les autres gènes) qui peut être bien estimé, car A et W sont continus et assez réguliers, alors que la relation entre Y et des centaines de gènes est difficile à apprendre avec $n = 38$.

— **Efficacité locale.** Si les deux nuisances sont estimées de façon convergente avec la condition de reste (ii), $\hat{\beta}$ atteint la borne d'efficacité du modèle partiellement linéaire homoscédastique. « Locale » signifie que l'efficacité est obtenue au point (sous-modèle) où les hypothèses de travail sont exactes.

— **Si un seul modèle est correct**, $\hat{\beta}$ reste asymptotiquement normal, mais sa courbe d'influence comporte un terme supplémentaire provenant de l'estimation de la nuisance correcte ; l'erreur type fondée sur D^* peut alors être légèrement erronée (souvent conservatrice). Avec de l'apprentissage machine, la condition de reste n'est pas garantie (van der Laan et Rose, 2011, encadré p. 96) ; il s'agit d'une hypothèse à garder en tête.

5.3 Inférence par la courbe d'influence

Puisque $\hat{\beta} - \beta_0 \approx P_n D^*$, la variance asymptotique est estimée par la variance empirique de la courbe d'influence estimée :

$$\widehat{\mathrm{ET}}(\hat{\beta}) = \frac{1}{\sqrt{n}} \left\{ \frac{1}{n-1} \sum_{i=1}^{n} (\mathrm{IC}_n(O_i) - \overline{\mathrm{IC}}_n)^2 \right\}^{1/2}, \qquad \mathrm{IC}_n(O) = \frac{H_n(A,W)\{Y - \hat{\beta}A - r_n^1(W)\}}{P_n H_n^2}.$$

Par (8), $\overline{\mathrm{IC}}_n = 0$. L'intervalle de confiance de Wald est $\hat{\beta} \pm z_{1-\alpha/2}\widehat{\mathrm{ET}}$ et le test de $H_0 : \beta = 0$ utilise $Z = \hat{\beta}/\widehat{\mathrm{ET}}$. On reconnaît la structure d'un estimateur sandwich : $P_n H_n^2$ est le « pain » et $P_n\{H_n^2(Y - \hat{\beta}A - r_n^1)^2\}$ la « viande ».

Attention

Avec $n = 38$, l'approximation normale et l'estimation de la variance par la courbe d'influence sont fragiles : les ajustements adaptatifs sur un petit échantillon tendent à sous-estimer les résidus, donc l'erreur type. Les valeurs p servent surtout à *classer* les gènes ; leur interprétation absolue demande de la prudence. Le préfiltrage des gènes par un test marginal sur les mêmes données (étape qui produit la liste de 681 gènes de l'étude de cas) est une double utilisation des données dont l'inférence ne tient pas compte.

5.4 Classer les gènes et contrôler la FDR

L'analyse se répète pour chaque gène candidat $A_1, \ldots, A_G$, chacun avec son propre ensemble d'ajustement W_g. On obtient des estimations $\hat{\beta}_g$, des statistiques Z_g et des valeurs p p_g. Le chapitre ordonne les gènes selon ces valeurs p (tableau 22.4) et les compare au classement marginal : des gènes très associés marginalement peuvent reculer une fois ajustés pour les gènes corrélés, et inversement. Pour déclarer une liste de gènes « significatifs », on contrôle le taux de fausses découvertes, par exemple par la procédure de Benjamini et Hochberg : $p_{(1)} \leq \cdots \leq p_{(G)}$, rejeter les k plus petites où $k = \max\{j : p_{(j)} \leq jq/G\}$. La justification de cette procédure, et son comportement sous dépendance entre les tests (les statistiques Z_g sont corrélées puisque les gènes le sont), sont traités à la section 11.

À retenir

- $\widehat{\mathrm{ET}} = \mathrm{sd}(\mathrm{IC}_n)/\sqrt{n}$ avec $\mathrm{IC}_n = H_n(Y - \hat{\beta}A - r_n^1)/P_n H_n^2$; test de Wald pour $\beta = 0$.
- Doublement robuste : convergent si g *ou* r est bien estimé ; localement efficace si les deux le sont, avec un reste $\|g_n - g_0\| \, \|r_n - r_0\| = o_P(n^{-1/2})$.
- Pour des milliers de gènes : une valeur p par gène, puis contrôle de la FDR (section 11).

6 Exemple en R sur des données simulées

Pour ne pas résoudre la question 3 de l'étude de cas (le gène TOP2B dans `golub`), on illustre l'algorithme sur des données simulées où la vraie valeur de β est connue. On simule $n = 300$ sujets avec cinq covariables $W_1, \ldots, W_5 \overset{\text{i.i.d.}}{\sim} \mathcal{U}(-1,1)$,

$$A = \underbrace{0{,}6\, W_1 + 0{,}4\, W_2^2}_{g_0(W)} + U, \quad U \sim \mathcal{U}(-1,1),$$

$$\mathbb{P}(Y=1 \mid A,W) = \underbrace{0{,}38 + 0{,}12\sin(\pi W_1) + 0{,}28\,W_2^2}_{r_0(W)} + 0{,}1\,A.$$

Ainsi $\beta_0 = 0{,}1$, et W_1, W_2 confondent la relation entre A et Y de façon non linéaire (le terme W_2^2 est orthogonal à W_2, de sorte qu'un ajustement linéaire ne le capte pas). Les probabilités restent dans $[0{,}10; 0{,}98]$.

```
library(earth)
genere <- function(n, beta = 0.1) {
  W <- matrix(runif(n * 5, -1, 1), n, 5, dimnames = list(NULL, paste0("W", 1:5)))
  gW <- 0.6 * W[, 1] + 0.4 * W[, 2]^2           # E[A | W]
  A  <- gW + runif(n, -1, 1)                     # biomarqueur
  rW <- 0.38 + 0.12 * sin(pi * W[, 1]) + 0.28 * W[, 2]^2
  Y  <- rbinom(n, 1, rW + beta * A)              # P(Y=1|A,W) = r(W) + beta*A
  data.frame(Y, A, W)
}
```

La fonction suivante implante les étapes 1 à 6 de la section 4. Les prédictions non paramétriques sont faites par `earth` (splines de régression adaptatives multivariées, MARS, avec interactions d'ordre 2), en version binomiale pour $\mathrm{E}[Y \mid A,W]$ (proportion) et gaussienne pour g et r (variables réelles).

```
tmle_lin <- function(d) {
  W <- d[, grep("^W", names(d))]
  # Étape 1 : estimateur initial de E[Y|A,W] (proportion -> binomiale)
  fitQ <- earth(x = cbind(A = d$A, W), y = d$Y, degree = 2,
                glm = list(family = binomial))
  Q_A <- predict(fitQ, newdata = cbind(A = d$A, W), type = "response")
  Q_0 <- predict(fitQ, newdata = cbind(A = 0, W), type = "response")
  beta0 <- coef(lm(I(Q_A - Q_0) ~ 0 + d$A))[[1]]   # projection sur beta*A
  # r_n^0(W) = E[Y - beta0 A | W] (variable réelle -> gaussienne)
  fitr <- earth(x = W, y = d$Y - beta0 * d$A, degree = 2)
  r0 <- predict(fitr, newdata = W)[, 1]
  # Étape 2 : mécanisme de "traitement" g(W) = E[A | W]
  fitg <- earth(x = W, y = d$A, degree = 2)
  g <- predict(fitg, newdata = W)[, 1]
  H <- d$A - g                                     # Étape 3 : covariable astucieuse
  # Étape 4 : fluctuation par moindres carrés avec décalage
  Q0 <- beta0 * d$A + r0
  eps <- coef(lm(d$Y ~ 0 + H, offset = Q0))[[1]]
  # Étape 5 : mise à jour et substitution
  beta1 <- beta0 + eps
  r1 <- r0 - eps * g
  # Étape 6 : courbe d'influence et inférence
  IC <- H * (d$Y - beta1 * d$A - r1) / mean(H^2)
  se <- sd(IC) / sqrt(nrow(d))
  c(beta0 = beta0, eps = eps, beta = beta1, se = se,
    z = beta1 / se, p = 2 * pnorm(-abs(beta1 / se)))
}
set.seed(7360)
d <- genere(300)
coef(summary(lm(Y ~ A, data = d)))["A", ]      # sans ajustement
coef(summary(lm(Y ~ ., data = d)))["A", ]      # ajustement linéaire pour W
tmle_lin(d)
```

```
  Estimate Std. Error    t value   Pr(>|t|)
    0.1368     0.0434     3.1524     0.0018
  Estimate Std. Error    t value   Pr(>|t|)
    0.1048     0.0511     2.0502     0.0412
 beta0     eps    beta      se       z       p
0.1062 -0.0097  0.0965  0.0507  1.9015  0.0572
```

Sur cet échantillon, l'analyse marginale surestime l'effet (0,137) ; l'ajustement linéaire et le TMLE donnent des valeurs proches de $\beta_0 = 0{,}1$. La correction $\varepsilon_n = -0{,}010$ est petite : l'estimateur initial était déjà raisonnable ici. Un seul échantillon ne dit rien du biais ni de la validité de l'erreur type ; pour cela, on répète l'expérience.

6.1 Vérification par simulation de l'erreur type

On répète la génération et l'estimation 1000 fois avec $n = 300$, puis 300 fois avec $n = 1000$ (code ci-dessous pour $n = 300$).

```
set.seed(2027)
res <- t(replicate(1000, {
  d <- genere(300)
  c(naif = coef(lm(Y ~ A, data = d))[["A"]],
    lin  = coef(lm(Y ~ ., data = d))[["A"]],
    tmle_lin(d)[c("beta0", "beta", "se")])
}))
colMeans(res[, c("naif", "lin", "beta0", "beta")]) - 0.1   # biais
c(sd(res[, "beta"]), mean(res[, "se"]))                     # ET empirique, ET moyenne
mean(abs(res[, "beta"] - 0.1) <= 1.96 * res[, "se"])        # couverture
```

```
   naif     lin   beta0    beta
 0.0689  0.0285 -0.0131 -0.0001
[1] 0.0526 0.0473
[1] 0.925
```

Le tableau 1 résume ces résultats et ceux de trois expériences complémentaires : $n = 1000$ (300 répétitions) ; une version avec *ajustement croisé* à 5 plis, où chaque prédiction de $\bar{Q}$, g et r est faite par un modèle ajusté sans le sujet concerné (500 répétitions) ; et un estimateur « oracle » qui utilise les vraies fonctions g_0 et r_0 (2000 répétitions).

Plusieurs leçons se dégagent (voir aussi la figure 1).

— L'analyse marginale est fortement biaisée par la confusion ; l'ajustement linéaire pour W ne retire qu'une partie du biais, car la confusion passe par W_2^2 et $\sin(\pi W_1)$.

— L'estimateur de substitution β_n^0 est biaisé (biais de régularisation de `earth`) : dans 23,5 % des répétitions, l'élagage de `earth` retire A du modèle et $\beta_n^0 = 0$ exactement (pic de la figure 1). L'algorithme a été réglé pour prédire Y, pas pour estimer β. L'étape de ciblage retire ce biais : le TMLE est pratiquement sans biais.

— L'erreur type par courbe d'influence estime correctement la variance *asymptotique* : elle coïncide avec l'écart type de l'oracle (0,047), qui atteint la borne d'efficacité. À $n = 300$, la variance réelle du TMLE dépasse toutefois cette borne d'environ 24 % (écart type supérieur de 11 %), parce que les termes de second ordre dus à l'estimation de g et de r ne sont pas encore

TABLE 1 – Simulation : biais, écart type empirique (ET emp.) et erreur type moyenne par courbe d'influence (ET IC) du TMLE de $\beta_0 = 0{,}1$, et couverture de l'intervalle de Wald à 95 %.

Estimateur	n	Biais	ET emp.	ET IC	Couverture
Marginal (`lm(Y ~ A)`)	300	0,069			
Linéaire ajusté pour W	300	0,029			
Initial β_n^0 (substitution)	300	−0,013			
TMLE	300	0,000	0,053	0,047	0,925
TMLE avec ajustement croisé	300	0,000	0,051	0,047	0,928
Oracle (g_0, r_0 connus)	300		0,047	0,047	
Initial β_n^0	1000	−0,008			
TMLE	1000	0,000	0,028	0,026	0,933

négligeables ; la couverture est de 92,5 %. L'écart diminue avec n (rapport des écarts types de 0,90 à $n = 300$, de 0,94 à $n = 1000$), comme le prévoit le théorème 5.1.

— L'ajustement croisé ne change presque rien ici : la sous-estimation ne vient pas d'un surajustement des prédictions, mais de la lenteur de convergence des nuisances. Avec les $n = 38$ échantillons de Golub, il faut donc interpréter les valeurs p avec prudence.

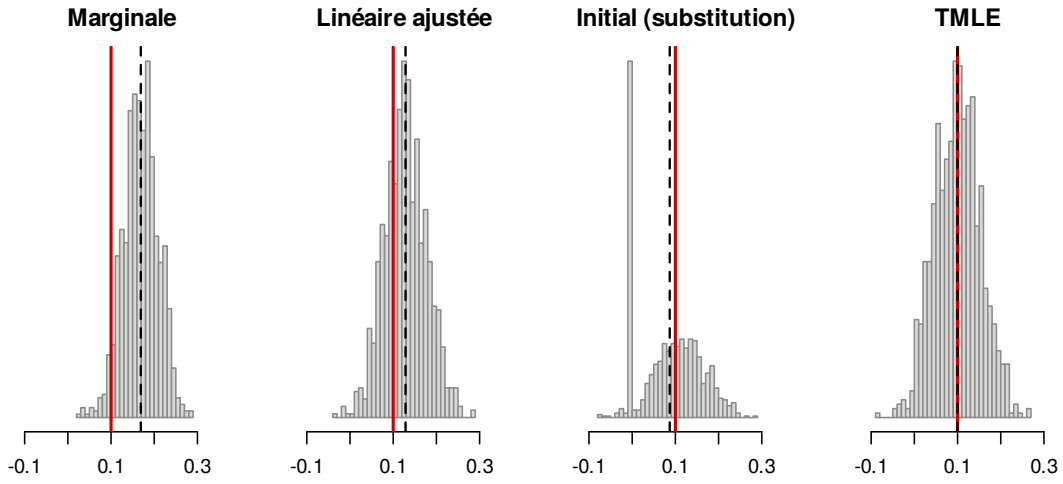


FIGURE 1 – Loi d'échantillonnage simulée (1000 répétitions, $n = 300$) de quatre estimateurs de $\beta_0 = 0{,}1$ (trait vertical) : régression marginale, régression linéaire ajustée pour W, estimateur initial de substitution β_n^0 et TMLE.

7 Niveau doctorat : courbe d'influence et modèle partiellement linéaire

Niveau doctorat – Score efficace du modèle partiellement linéaire

Modèle. $Y = \beta A + r(W) + \epsilon$ avec $\mathrm{E}[\epsilon \mid A,W] = 0$; les nuisances sont r, la loi $p(W,A)$ et la loi conditionnelle de ϵ sachant (A,W) (sous la seule contrainte de moyenne nulle). C'est un modèle de moments restreints (Tsiatis, 2006, § 4.5) dont la moyenne $\mu(A,W) = \beta A + r(W)$ contient elle-même une nuisance de dimension infinie.
Étape 1 : modèle de moments restreints. Pour r fixée, Tsiatis montre que l'orthogonal de l'espace tangent de nuisance est $\Lambda^{\perp} = \{\alpha(A,W)\,\epsilon : \alpha \text{ quelconque}\}$, et que le score efficace pour un paramètre de la moyenne est $\{\partial\mu/\partial\theta\}\,\sigma^{-2}(A,W)\,\epsilon$.
Étape 2 : ajouter la nuisance r. Un sous-modèle paramétrique $r_t = r_0 + t\,s(W)$ a pour score, projeté sur $\Lambda^{\perp}$, la fonction $s(W)\sigma^{-2}(A,W)\epsilon$. Le score efficace pour β est le résidu de la projection de $A\sigma^{-2}\epsilon$ sur l'adhérence de $\{s(W)\sigma^{-2}\epsilon\}$: on cherche s^* tel que $\mathrm{E}[\{A - s^*(W)\}\sigma^{-2}\epsilon \cdot s(W)\sigma^{-2}\epsilon] = 0$ pour tout s, soit $\mathrm{E}[\{A - s^*(W)\}\sigma^{-2}(A,W) \mid W] = 0$, d'où

$$s^*(W) = \frac{\mathrm{E}[A\,\sigma^{-2}(A,W) \mid W]}{\mathrm{E}[\sigma^{-2}(A,W) \mid W]}, \qquad S_{\mathrm{eff}}(O) = \frac{A - s^*(W)}{\sigma^2(A,W)}\,\{Y - \beta A - r(W)\}.$$

Étape 3 : cas homoscédastique. Si $\sigma^2(A,W) \equiv \sigma^2$, alors $s^* = g_0$ et $S_{\mathrm{eff}} = \sigma^{-2}\{A - g_0(W)\}\{Y - \beta A - r(W)\}$. On retrouve le score efficace $\mathrm{E}[(A - \mathrm{E}[A \mid W])(Y - \beta A - \theta(W))]$ (avec $\theta = r$) donné en équation estimante. La courbe d'influence efficace est $D^* = S_{\mathrm{eff}} / \mathrm{E}[S_{\mathrm{eff}}^2]$; comme $\mathrm{E}[S_{\mathrm{eff}}^2] = \sigma^{-2}\,\mathrm{E}[\{A - g_0\}^2]$, le facteur σ^2 se simplifie et l'on obtient (5). Pour Y dichotomique, $\sigma^2 = \bar{Q}(1 - \bar{Q})$ n'est pas constante : l'estimateur efficace pondérerait par $1/\{\bar{Q}(1 - \bar{Q})\}$ et remplacerait g_0 par s^* ; le TMLE du chapitre utilise l'hypothèse de travail homoscédastique.

Niveau doctorat – Courbe d'influence du TMLE (défi H26)

Développement. Notons $\phi(\beta, r, g)(O) = \{A - g(W)\}\{Y - \beta A - r(W)\}$, $H_0 = A - g_0(W)$, $\hat{\phi} = \phi(\beta_0, r_n^1, g_n)$ et $\phi_0 = \phi(\beta_0, r_0, g_0)$. Comme ϕ est linéaire en β, l'équation résolue (8) s'écrit

$$0 = P_n\phi(\hat{\beta}, r_n^1, g_n) = P_n\hat{\phi} - (\hat{\beta} - \beta_0)\,P_n\{(A - g_n)A\}.$$

Donc

$$\hat{\beta} - \beta_0 = \frac{P_n\hat{\phi}}{P_n\{(A - g_n)A\}}, \qquad P_n\hat{\phi} = \underbrace{(P_n - P_0)\phi_0}_{(a)} + \underbrace{(P_n - P_0)(\hat{\phi} - \phi_0)}_{(b)} + \underbrace{P_0\hat{\phi}}_{(c)}.$$

- (a) est une moyenne de variables i.i.d. centrées : c'est le terme principal.
- $(b) = o_P(n^{-1/2})$ si $\|\hat{\phi} - \phi_0\|_2 \xrightarrow{\mathbb{P}} 0$ et si $\hat{\phi}$ appartient à une classe de Donsker (lemme d'équicontinuité, van der Vaart, 1998, ch. 19) ; l'ajustement croisé (estimer g_n, r_n sur une partie des données, évaluer ϕ sur l'autre) permet de se passer de la condition de Donsker.
- $(c) = P_0[(g_0 - g_n)(r_0 - r_n^1)]$ par la proposition 3.2 ; par Cauchy-Schwarz, $|(c)| \le \|g_n - g_0\|_2 \|r_n^1 - r_0\|_2 = o_P(n^{-1/2})$.

Enfin, $P_n\{(A-g_n)A\} \xrightarrow{\mathbb{P}} \mathrm{E}_0[H_0 A] = \mathrm{E}_0[H_0^2]$, puisque $\mathrm{E}_0[H_0\, g_0(W)] = 0$. Par le lemme de Slutsky,

$$\hat{\beta} - \beta_0 = \frac{1}{n}\sum_{i=1}^{n} \frac{H_0(A_i,W_i)\{Y_i - \beta_0 A_i - r_0(W_i)\}}{\mathrm{E}_0[H_0^2]} + o_P(n^{-1/2}) = P_n D^* + o_P(n^{-1/2}).$$

Remarque. Si $g_n \to g_0$ mais $r_n^1 \to r_\infty \neq r_0$, le terme (c) n'est plus négligeable a priori : $(c) = P_0[(g_0 - g_n)(r_0 - r_\infty)] + \ldots$ est d'ordre $\|g_n - g_0\|$, et sa contribution linéaire dépend de l'estimateur g_n. La courbe d'influence n'est alors plus D^* (sauf si g_n est un estimateur paramétrique efficace, cas où l'on peut montrer que le terme supplémentaire réduit la variance).

Niveau doctorat – Interprétation non paramétrique de la même courbe d'influence

Soit $\beta^*(P) = N(P)/D(P)$ avec $N = \mathrm{E}[\mathrm{Cov}(A,Y \mid W)] = \mathrm{E}[AY] - \mathrm{E}[g(W)m(W)]$, $m(W) = \mathrm{E}[Y \mid W]$, et $D = \mathrm{E}[\mathrm{Var}(A \mid W)] = \mathrm{E}[\{A - g(W)\}^2]$. Dans le modèle *non paramétrique*, les fonctions d'influence (dérivées de Gâteaux le long de contaminations ponctuelles) sont :

$$\mathrm{IF}_{\mathrm{E}[AY]} = AY - \mathrm{E}[AY], \qquad \mathrm{IF}_{\mathrm{E}[gm]} = gm - \mathrm{E}[gm] + g(W)\{Y - m(W)\} + m(W)\{A - g(W)\},$$

d'où $\mathrm{IF}_N = \{A-g(W)\}\{Y-m(W)\} - N$ et, de même, $\mathrm{IF}_D = \{A-g(W)\}^2 - D$. Par la règle du quotient, et puisque $N - \beta^* D = 0$,

$$\mathrm{IF}_{\beta^*} = \frac{\mathrm{IF}_N - \beta^*\,\mathrm{IF}_D}{D} = \frac{\{A-g(W)\}\{Y - \beta^* A - (m(W) - \beta^* g(W))\}}{\mathrm{E}[\{A-g(W)\}^2]}.$$

Avec $r(W) = m(W) - \beta^* g(W)$, c'est exactement (5). Comme le modèle non paramétrique n'a qu'un gradient, c'est la courbe d'influence *efficace* de β^* dans ce modèle. Conséquence pratique : l'erreur type $\mathrm{sd}(\mathrm{IC}_n)/\sqrt{n}$ reste valide pour β^* même si l'effet de A n'est pas linéaire, à condition que r_n^1 estime $m - \beta^* g$, ce que fait la construction $r_n^0 = \hat{\mathrm{E}}[Y - \beta_n^0 A \mid W]$ suivie de la mise à jour. Vansteelandt et Dukes (2022) développent cette idée d'« inférence sans hypothèse forte » pour les paramètres de modèles linéaires généralisés.

8 Préparation de l'étude de cas

Les consignes détaillées et les questions sont dans le document de consignes de l'étude de cas 2. Quelques repères pour la lecture du chapitre :

— la sous-section sur le paramètre d'importance de variable et le modèle semi-paramétrique correspond à notre section 2 ;

— les deux encadrés gris de la sous-section *Implementation* (p. 370–371) correspondent à notre algorithme (section 4) ;

— la sous-section *Statistical inference* (bas de la p. 371) donne la courbe d'influence (5) utilisée pour l'erreur type ;

— l'application aux données de Golub (section 22.5) et le tableau 22.4 présentent le classement des gènes ; le chapitre utilise des splines polynomiales pour les prédictions non paramétriques

(p. 378).

Pour les données `golub` du module `multtest` : la matrice `golub` (3051×38) contient les expressions, `golub.cl` la classe ($0 = \text{ALL}, 1 = \text{AML}$) et `golub.gnames` les noms des gènes. Attention à l'orientation de la matrice (gènes en lignes) et au codage de Y, qui détermine le signe de $\hat{\beta}$.

9 Synthèse

Synthèse après la séance – Tuglus et van der Laan (2011)

Objet d'intérêt	Pour chaque gène A : l'effet marginal ajusté $\psi(a) = \mathrm{E}_W[\mathrm{E}(Y \mid A = a, W) - \mathrm{E}(Y \mid A = 0, W)] = \beta a$, résumé par la pente β du modèle semi-paramétrique $\mathrm{E}[Y \mid A, W] = \beta A + r(W)$ (ou $\beta^* = \mathrm{E}[\mathrm{Cov}(A, Y \mid W)] / \mathrm{E}[\mathrm{Var}(A \mid W)]$ hors modèle) ; ensuite, une liste ordonnée de gènes.
Nuisance / biais	Confusion par les autres gènes W (corrélés à A et à Y) ; fonctions nuisibles de grande dimension $r(W)$ (ou $\mathrm{E}[Y \mid A, W]$) et $g(W) = \mathrm{E}[A \mid W]$; biais de régularisation d'un estimateur de substitution ; colinéarité (positivité).
Approche	Modèle semi-paramétrique : aucune forme imposée pour r et g, estimées par apprentissage machine (splines adaptatives) ; restriction de W aux gènes peu corrélés à A ; ciblage par fluctuation dans la direction $H = A - g(W)$.
Critère	Perte quadratique (moindres carrés) pour $\bar{Q}$ et pour la fluctuation ; équivaut à résoudre l'équation estimante efficace $\sum_i \{A_i - g_n(W_i)\}\{Y_i - \beta A_i - r_n(W_i)\} = 0$.
Calcul	Ajustements non paramétriques (`polspline`/`earth`) de $\bar{Q}$, r et g ; régression linéaire sans ordonnée à l'origine de Y sur H_n avec décalage $\bar{Q}_n^0$; mise à jour en une étape $\hat{\beta} = \beta_n^0 + \varepsilon_n$.
Inférence	Normalité asymptotique ; erreur type $\mathrm{sd}(\mathrm{IC}_n)/\sqrt{n}$ par la courbe d'influence $\mathrm{IC} = H(Y - \beta A - r) / \mathrm{E}[H^2]$; test de Wald par gène ; double robustesse, efficacité locale ; contrôle de la FDR pour la liste de gènes.

Références

— Benjamini, Y. et Hochberg, Y. (1995). Controlling the false discovery rate : a practical and powerful approach to multiple testing. *JRSS B*, 57, 289–300.

— Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. et Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21, C1–C68.

— Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, 19, 1–67.

— Golub, T. R. et coll. (1999). Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, 286, 531–537.

— Kooperberg, C., Bose, S. et Stone, C. J. (1997). Polychotomous regression. *JASA*, 92, 117–127.

— Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56, 931–954.

— Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer (ch. 3 et 4).

— Tuglus, C. et van der Laan, M. J. (2011). Targeted methods for biomarker discovery. Dans van der Laan et Rose (2011), ch. 22.

— van der Laan, M. J. (2006). Statistical inference for variable importance. *International Journal of Biostatistics*, 2(1), art. 2.

— van der Laan, M. J. et Rose, S. (2011). *Targeted Learning : Causal Inference for Observational and Experimental Data*. Springer (ch. 4, 5 et annexe A1).

— van der Laan, M. J. et Rubin, D. (2006). Targeted maximum likelihood learning. *International Journal of Biostatistics*, 2(1), art. 11.

— van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

— Vansteelandt, S. et Dukes, O. (2022). Assumption-lean inference for generalised linear model parameters. *JRSS B*, 84, 657–685.