

Efficacité asymptotique de quelques indices de sensibilité

T. Klein, A. Lagnoux et P. Rochet

Journées ANR GATSBII, 31 janvier 2025



Efficacité asymptotique

Cadre théorique :

- $\mathcal{P}$ un modèle statistique
- $Z_1, \dots, Z_n$ un échantillon iid de loi $P \in \mathcal{P}$ inconnue
- $\|\cdot\|$ et $\langle \cdot, \cdot \rangle$ les norme et produit scalaire de $L^2(P)$

Définition

Un sous-modèle $(P_t)_{t \geq 0} \subseteq \mathcal{P}$ avec $P_0 = P$ est **différentiable en moyenne quadratique** en 0 s'il existe $g \in L^2(P)$ telle que

$$\lim_{t \rightarrow 0^+} \left\| \frac{\sqrt{f_t} - 1}{t} - \frac{1}{2}g \right\| = 0$$

où f_t est la densité de Radon-Nicodym de la partie absolument continue de P_t par rapport à P .

- Dérivabilité (Fréchet) dans $L^2(P)$ de $t \mapsto \sqrt{f_t}$ en 0 avec $f_0 = 1$, P -ps
- g est appelée *fonction de score*
- L'adhérence de l'espace vectoriel engendré par les fonctions de score est appelé l'*espace tangent* au modèle en P , et noté $\dot{\mathcal{P}}_P$.

Inférence :

- $\psi : \mathcal{P} \rightarrow \mathbb{R}$ un paramètre réel
- $\hat{\psi}_n = \hat{\psi}_n(Z_1, \dots, Z_n)$ un estimateur de $\psi(P)$

Définition

On appelle *fonction d'influence* de ψ en $P \in \mathcal{P}$, toute fonction $h \in L^2(P)$ telle que

$$\lim_{t \rightarrow 0^+} \frac{\psi(P_t) - \psi(P)}{t} = \langle h, g \rangle$$

pour tout $g \in \dot{\mathcal{P}}_P$ et sous-modèle différentiable $(P_t)_{t>0}$ de score g .

- Représentation de Riesz de la composition des dérivées en 0 dans $L^2(P)$
- Toute fonction d'influence se décompose de manière unique en

$$h = \tilde{\psi} + h^\perp$$

avec $\tilde{\psi} \in \dot{\mathcal{P}}_P$ (fonction d'influence *efficace*) et $h^\perp \in \dot{\mathcal{P}}_P^\perp$.

Définition

L'estimateur $\hat{\psi}_n$ est dit régulier s'il existe une loi L telle que

$$\sqrt{n}(\hat{\psi}_n - \psi(P_{1/\sqrt{n}})) \xrightarrow{\text{loi}} L$$

pour tout sous-modèle différentiable $\{P_t, t \in [0, \varepsilon]\}$, où pour tout $n \in \mathbb{N}$, l'échantillon $Z_1, \dots, Z_n$ est de loi $P_{1/\sqrt{n}}$.

Si $\hat{\psi}_n$ est un estimateur régulier et asymptotiquement gaussien

$$\sqrt{n}(\hat{\psi}_n - \psi(P)) \xrightarrow{\text{loi}} \mathcal{N}(0, \sigma^2)$$

alors $\sigma^2 \geq \|\tilde{\psi}\|^2$

Lemme (van der Vaart)

Un estimateur régulier $\hat{\psi}_n$ est asymptotiquement efficace si et seulement si

$$\hat{\psi}_n = \psi(P) + \frac{1}{n} \sum_{i=1}^n \tilde{\psi}(Z_i) + o_P(1/\sqrt{n})$$

Pick-Freeze

- **Cadre :**

- (Y_i, Y'_i) iid, Y_i, Y'_i indépendants de même loi conditionnellement à X_i
- P désigne ici la loi du couple (Y_i, Y'_i)

- **Modèle :** (Lois échangeables)

$$\mathcal{P} = \{P : \forall A, B \in \mathcal{B}(\mathbb{R}), P(A \times B) = P(B \times A)\}$$

- **Espace tangent :**

$$\dot{\mathcal{P}}_P = \{g \in L^2(P) : \langle g, 1 \rangle = 0, g(y, y') = g(y', y)\}$$

Référence

1. A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. *Asymptotic normality and efficiency of two Sobol index estimators*. ESAIM: Probability and Statistics, 2014

- Paramètre :

$$\psi(P) = \mathbb{E}[\mathbb{E}(Y|X)^2] = \mathbb{E}[YY'] \quad , \quad (Y, Y') \sim P$$

- Calcul de la fonction d'influence efficace :

- Soit $P_t = (1 + tg)P$ de score g , (Y_t, Y'_t) de loi P_t
- $\psi(P_t) = \mathbb{E}[Y_t Y'_t] = \mathbb{E}[(1 + tg(Y, Y')) YY']$
- $\frac{\psi(P_t) - \psi(P)}{t} = \mathbb{E}[g(Y, Y') YY'] = \langle h, g \rangle$ où $h(y, y') = yy'$
- $\tilde{\psi} : (y, y') \mapsto yy' - \psi(P) \in \dot{\mathcal{P}}_P$ (symétrique, centrée)

- Estimation :

$$\widehat{\psi}_n = \frac{1}{n} \sum_{i=1}^n Y_i Y'_i = \psi(P) + \frac{1}{n} \sum_{i=1}^n \tilde{\psi}(Y_i, Y'_i)$$

est asymptotiquement efficace.

- **Paramètre :**

- $\forall z \in \mathbb{R}, \psi_z(P) = \mathbb{E}[\mathbb{P}(Y \leq z|X)^2] = \mathbb{E}[\mathbb{1}\{Y \leq z, Y' \leq z\}]$
- $\psi^Q(P) = \int \psi_z(P) dQ(z)$, pour Q une mesure de proba (connue)

- **Calcul de la fonction d'influence efficace :**

- ψ_z a pour fonction d'influence efficace

$$\tilde{\psi}_z(y, y') = \mathbb{1}\{y \leq z, y' \leq z\} - \psi_z(P)$$

- Par Fubini

$$\frac{d}{dt} \psi^Q(P_t) \Big|_{t=0} = \int \left(\frac{d}{dt} \psi_z(P_t) \right) \Big|_{t=0} dQ(z) = \int \langle \tilde{\psi}_z, g \rangle dQ(z) = \langle \tilde{\psi}^Q, g \rangle$$

avec

$$\tilde{\psi}^Q(y, y') = \int \tilde{\psi}_z(y, y') dQ(z) = Q([y \vee y'; \infty]) - \psi^Q(P)$$

- **Estimation :**

$$\hat{\psi}_n = \int \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{Y_i \leq z, Y'_i \leq z\} dQ(z) = \frac{1}{n} \sum_{i=1}^n Q([Y_i \vee Y'_i; \infty])$$

est asymptotiquement efficace.

Si Q est inconnue et on observe $Z_1, \dots, Z_n$ iid de loi Q (indépendants des Y_i, Y'_i)

- **Modèle :**

$$\mathcal{P} \otimes \mathcal{Q} := \{P \otimes Q : P \in \mathcal{P}, Q \in \mathcal{Q}\}$$

- **Espace tangent :**

$$(\mathcal{P} \dot{\otimes} \mathcal{Q})_{P \otimes Q} = \dot{\mathcal{P}}_P \oplus \dot{\mathcal{Q}}_Q$$

Lemme

Fonction d'influence efficace du paramètre $\Psi(P \otimes Q) = \psi^Q(P)$, $P \in \mathcal{P}, Q \in \mathcal{Q}$:

$$\tilde{\Psi}(y, y', z) = \tilde{\psi}^Q(y, y') + \psi_z(P) - \psi^Q(P)$$

- **Estimation :**

$$\hat{\psi}_n = \frac{1}{n^2} \sum_{i,j=1}^n \mathbb{1}\{Y_i \leq Z_j, Y'_i \leq Z_j\}$$

asymptotiquement efficace?

Si l'intégrale se fait contre la loi de Y (pas de données supplémentaires Z_i)

- **Modèle :** (Lois échangeables)

$$\mathcal{P} = \{P : \forall A, B \in \mathcal{B}(\mathbb{R}), P(A \times B) = P(B \times A)\}$$

- **Paramètre :**

$$\Psi(P) = \psi^P(P) = \int \psi_z(P) dP(z, z')$$

Lemme

Fonction d'influence efficace du paramètre Ψ en P :

$$\tilde{\Psi}(y, y') = \tilde{\psi}^P(y, y') + \frac{\psi_y(P) + \psi_{y'}(P)}{2} - \psi^P(P)$$

- Calcul de la fonction d'influence efficace :

$$\begin{aligned}
 - \quad \Psi(P_t) &= \int \psi_z(P_t) dP_t(z, z') \\
 &= \int \psi_z(P_t) dP(z, z') + t \int \psi_z(P_t) g(z, z') dP(z, z') + o(t) \\
 &= \Psi(P) + t \int (\tilde{\psi}^P(z, z') + \psi_z(P)) g(z, z') dP(z, z') + o(t) \\
 &= \Psi(P) + t \langle \tilde{\psi}^P + \psi_{\cdot}(P), g \rangle + o(t)
 \end{aligned}$$

- La fonction

$$(y, y') \mapsto \tilde{\psi}^P(y, y') + \psi_y(P)$$

est une fonction d'influence mais n'appartient pas à $\dot{\mathcal{P}}_P$ (pas efficace).

- Il ne reste qu'à symétriser

$$(y, y') \mapsto \tilde{\psi}^P(y, y') + \frac{\psi_y(P) + \psi_{y'}(P)}{2} - \psi^P(P) \in \dot{\mathcal{P}}_P$$

- Estimation :

$$\hat{\psi}_n = ??$$

Given data

- **Données :**

$$(X_i, Y_i), i = 1, \dots, n \text{ iid de loi } P$$

- **Modèle non-paramétrique :**

$$\mathcal{P} = \mathcal{P}(\mathbb{R}^k \times \mathbb{R})$$

- **Espace tangent maximal :**

$$\dot{\mathcal{P}}_P = \{g \in L^2(P) : \langle g, 1 \rangle = 0\}$$

- **Paramètre :** Posons $m(X) = \mathbb{E}(Y|X)$

$$\psi(P) = \mathbb{E}[m(X)^2] = \mathbb{E}[Ym(X)]$$

- Soit $P_t = (1 + tg)P$, (X_t, Y_t) de loi P_t et $m_t(x) = \mathbb{E}[Y_t|X_t = x]$

-

$$\begin{aligned}\mathbb{E}[Y_t m(X_t)] &= \mathbb{E}[m_t(X_t) m(X_t)] \\ \iff \mathbb{E}[Y m(X)(1 + tg(X, Y))] &= \mathbb{E}[m_t(X) m(X)(1 + tg(X, Y))] \\ \iff \mathbb{E}\left[\frac{m_t(X) - m(X)}{t} Y\right] &= \mathbb{E}[(Y - m_t(X)) m(X) g(X, Y)]\end{aligned}$$

-

$$\begin{aligned}\psi(P_t) &= \mathbb{E}[Y_t m_t(X_t)] \\ &= \mathbb{E}[Y m_t(X)] + t \mathbb{E}[g(X, Y) Y m_t(X)] \\ &= \psi(P) + t \mathbb{E}[(m_t(X) - m(X)) Y] + t \mathbb{E}[g(X, Y) Y m_t(X)] \\ &= \psi(P) + t \mathbb{E}[(Y - m_t(X)) m(X) + Y m_t(X)] g(X, Y)\end{aligned}$$

Lemme

Si g est bornée P presque sûrement,

$$\lim_{t \rightarrow 0^+} \|m_t - m\| = 0$$

- **Fonction d'influence efficace :**

$$\tilde{\psi}(x, y) = 2y m(x) - m^2(x) - \psi(P)$$

- **Estimation efficace :**

- Combinaison d'estimateurs des rangs (en dimension 1)
- Noyaux (toute dimension)
- k -NN pondérés ?

Références

2. Doksum et Samarov, *Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression*, The Annals of Statistics, 1995
3. Da Veiga and Gamboa, *Efficient estimation of sensitivity indices*, Journal of Nonparametric Statistics, 2013
4. Klein et Rochet, *Efficiency of the averaged rank-based estimator for first order Sobol index inference*, Statistics & Probability Letters, 2024
5. Da Veiga, Gamboa, Lagnoux, Klein et Prieur, *Efficient estimation of Sobol' indices of any order from a single input/output sample*, Preprint

- **Paramètre :** Posons $p_z(x) = \mathbb{P}(Y \leq z | X = x)$

- $\forall z \in \mathbb{R}, \psi_z(P) = \mathbb{E}[p_z(X)^2]$

- $\psi^Q(P) = \int \psi_z(P) dQ(z)$, pour Q une mesure de proba (connue)

- **Calcul de la fonction d'influence efficace :**

- ψ_z a pour fonction d'influence efficace

$$\tilde{\psi}_z(x, y) = 2\mathbb{1}\{y \leq z\} p_z(x) - p_z^2(x) - \psi_z(P)$$

- Toujours par Fubini ...

$$\tilde{\psi}^Q(x, y) = \int \tilde{\psi}_z(x, y) dQ(z)$$

- **Estimation :**

- $\hat{\psi}_n(z)$ asymptotiquement efficace à z fixé

- Sous les bonnes hypothèses

$$\hat{\psi}_n = \int \hat{\psi}_n(z) dQ(z)$$

devrait rester asymptotiquement efficace...

Reste à traiter

- Q inconnue et échantillon $Z_1, \dots, Z_n$ iid de loi Q
- $Q = P_Y$ (Cramér - von Mises)
- Estimateur asymptotiquement efficace (Pick-Freeze et Given data)