

Algorithmes de Newton stochastiques

A. Godichon-Baggioni

Algorithme de Newton stochastique

IDÉE

Modèle linéaire :

$$Y = \theta^T X + \epsilon$$

Gradient stochastique : Posons $H = \mathbb{E} [XX^T] = \begin{pmatrix} 10^{-2} & 0 \\ 0 & 10^2 \end{pmatrix}$.

$$\mathbb{E} [\theta_{n+1} - \theta] = \begin{pmatrix} 1 - \gamma_{n+1}10^{-2} & 0 \\ 0 & 1 - \gamma_{n+1}10^2 \end{pmatrix} \mathbb{E} [\theta_n - \theta]$$

Choix de γ_n ?

SIMULATIONS

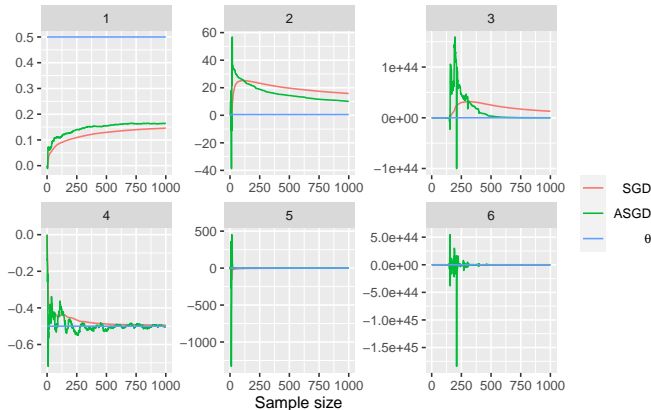


FIGURE – Evolution des estimateurs de la première coordonnée (en haut) et de la deuxième (en bas) pour, de gauche à droite, $c_\gamma = 0.1$, $c_\gamma = 1$ et $c_\gamma = 10$.

SIMULATIONS (NEWTON STOCHASTIQUE)

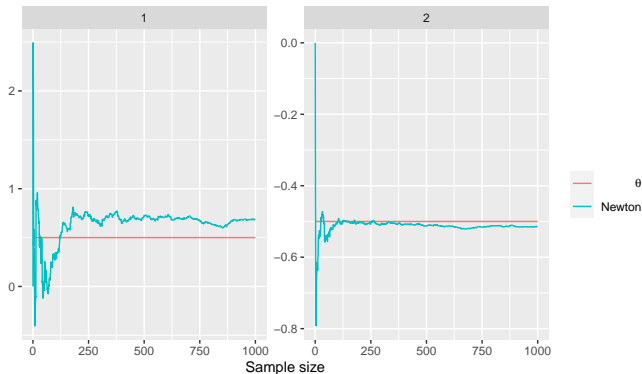


FIGURE – Evolution des estimateurs de la première coordonnée (à gauche) et de la deuxième coordonnée (à droite) pour l'algorithme de Newton stochastique.

ALGORITHME DE NEWTON STOCHASTIQUE

Algorithme de Newton stochastique :

$$m_{n+1} = m_n - \frac{1}{n+1} \bar{H}_n^{-1} \nabla_h g(X_{n+1}, m_n)$$

Hypothèses sur \bar{H}_n :

- ▶ \bar{H}_n^{-1} est symétrique et définie positive.
- ▶ Il existe une filtration (\mathcal{F}_n) telle que
 - ▶ \bar{H}_n et m_n sont $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$ mesurables.
 - ▶ X_{n+1} est indépendant de \mathcal{F}_n .

Vitesse de convergence

CADRE

Hypothèses sur la fonction G :

(PS0'') Il existe une constante C telle que

$$\forall h \in \mathbb{R}^d, \quad \mathbb{E} \left[\|\nabla_h g(X, h)\|^2 \right] \leq C + C (G(h) - G(m))$$

(PS2) La fonction G est deux fois continument différentiable sur une voisinage de m et

$$\lambda_{\min} := \lambda_{\min} (\nabla^2 G(m)) > 0.$$

(PS5) La Hessienne de G est uniformément bornée : il existe $L_{\nabla G}$ tel que

$$\forall h \in \mathbb{R}^d, \quad \|\nabla^2 G(h)\|_{op} \leq L_{\nabla G}$$

- ▶ **(PS5)** $\implies \nabla G(\cdot)$ est $L_{\nabla G}$ Lipchitz.
- ▶ G fortement convexe + **(PS0)** \implies **(PS0'')**.
- ▶ **(PS0'')** + **(PS5)** \implies **(PS0)**.

HYPOTHÈSE SUR L'ESTIMATEUR \overline{H}_n

(H1) On peut contrôler les valeurs propres de \overline{H}_n :

$$\lambda_{\max}(\overline{H}_n) = O(1) \quad p.s$$

$$\lambda_{\max}(\overline{H}_n^{-1}) = O(n^\beta) \quad p.s$$

avec $\beta < 1/2$.

► **(H1)** $\implies \liminf \lambda_{\min}(\overline{H}_n^{-1}) > 0 \text{ p.s.}$

CONVERGENCE

Théorème

*On suppose que les hypothèses **(PS0'')**, **(PS2)**, **(PS5)** et **(H1)** sont vérifiées. Alors*

$$m_n \xrightarrow[n \rightarrow +\infty]{p.s.} m.$$

NOUVELLE HYPOTHÈSE SUR \bar{H}_n

(H2) Si **(PS0'')**, **(PS2)**, **(PS5)** et **(H1)** sont vérifiées, alors

$$\bar{H}_n \xrightarrow[n \rightarrow +\infty]{p.s.} H \quad \text{et} \quad \bar{H}_n^{-1} \xrightarrow[n \rightarrow +\infty]{p.s.} H^{-1}$$

$$\blacktriangleright m_n \xrightarrow[n \rightarrow +\infty]{p.s.} m \implies \bar{H}_n \xrightarrow[n \rightarrow +\infty]{p.s.} H.$$

VITESSE DE CONVERGENCE

Théorème

On suppose que les hypothèses (PS0'') (PS2), (PS5), (H1) et (H2) sont vérifiées. Alors, pour tout $\delta > 0$,

$$\|m_n - m\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \quad p.s.$$

EFFICACITÉ ASYMPTOTIQUE

(H3) On suppose que les hypothèses **(PS0'')** **(PS2)**, **(PS5)**, **(H1)** et **(H2)** sont vérifiées, alors il existe $p_H > 0$ tel que

$$\|\bar{H}_n - H\|_{op} = O\left(\frac{1}{n^{p_H}}\right) \quad p.s$$

$$\|\bar{H}_n^{-1} - H^{-1}\|_{op} = O\left(\frac{1}{n^{p_H}}\right) \quad p.s$$

- Avoir une vitesse pour m_n implique d'avoir une vitesse pour \bar{H}_n .

EFFICACITÉ ASYMPTOTIQUE

Théorème

On suppose que les hypothèses (PS0''), (PS2) à (PS5), et (H1) à (H3) sont vérifiées, alors

$$\sqrt{n} (m_n - m) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N} (0, H^{-1} \Sigma H^{-1})$$

avec $H = \nabla^2 G(m)$ et $\Sigma = \Sigma(m)$.

Régression linéaire

UNE FORMULE MAGIQUE

Formule de Riccati : Soit $A \in \mathcal{M}_d(\mathbb{R})$ une matrice inversible et $u, v \in \mathbb{R}^d$. Si $1 + v^T A^{-1} u \neq 0$, alors $A + uv^T$ est inversible et

$$(A + uv^T)^{-1} = A^{-1} - (1 + v^T A^{-1} u)^{-1} A^{-1} uv^T A^{-1}.$$

Cas particulier : Soit A une matrice définie positive, pour tout $u \in \mathbb{R}^d$ et $\lambda \geq 0$, on a $1 + \lambda u^T A^{-1} u \geq 1$ et donc

$$(A + \lambda uu^T)^{-1} = A^{-1} - \lambda (1 + \lambda u^T A^{-1} u)^{-1} A^{-1} uu^T A^{-1}.$$

L'ALGORITHME

Algorithme de Newton stochastique :

$$\theta_{n+1} = \theta_n + \frac{1}{n+1} \bar{H}_n^{-1} (Y_{n+1} - X_{n+1}^T \theta_n) X_{n+1}$$

$$H_{n+1}^{-1} = H_n^{-1} + (1 + X_{n+1}^T H_n^{-1} X_{n+1})^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}$$

avec H_0 positive et $\bar{H}_n = (n+1)H_n^{-1}$.

Réécriture de \bar{H}_n :

$$\bar{H}_n = \frac{1}{n+1} \left(H_0 + \sum_{k=1}^n X_k X_k^T \right).$$

VITESSE DE CONVERGENCE

Théorème

On suppose qu'il existe $\eta > 0$ tel que X et ϵ admettent des moments d'ordre $4 + \eta$ et $2 + \eta$. Alors pour tout $\delta > 0$,

$$\|\theta_n - \theta\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) p.s. \quad \text{et} \quad \sqrt{n}(\theta_n - \theta) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2 H^{-1})$$

SIMULATIONS

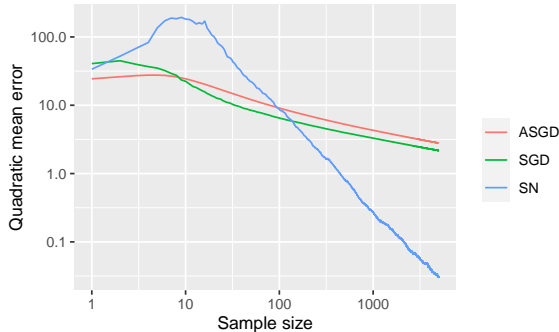


FIGURE – Evolution de l'erreur quadratique moyenne des estimateurs de gradient θ_n (SGD), de leur version moyennée $\bar{\theta}_n$ (ASGD) et des estimateurs de Newton stochastique $\tilde{\theta}_n$ (SN) en fonction de la taille de l'échantillon dans le cadre du modèle linéaire.

TESTER $H_0 : \theta = \theta_0$ "EN LIGNE"

Réécriture du TLC : Sous H_0 ,

$$\sqrt{n} \frac{(\theta_n - \theta_0)^T H (\theta_n - \theta_0)}{\sigma^2} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi_d^2$$

Application : Soit \bar{H}_n et $\hat{\sigma}_n^2$ des estimateurs consistants. Alors

$$K_n := \sqrt{n} \frac{(\theta_n - \theta_0)^T \bar{H}_n (\theta_n - \theta_0)}{\hat{\sigma}_n^2} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi_d^2$$

CONSTRUCTION DE \bar{H}_n ET $\hat{\sigma}_n^2$

Ecriture directe :

$$\bar{H}_n = \frac{1}{n+1} \left(H_0 + \sum_{k=1}^n X_k X_k^T \right)$$

$$\hat{\sigma}_n^2 = \frac{1}{n+1} \sum_{k=1}^n (Y_k - X_k^T \theta_{k-1})^2$$

Ecriture récursive :

$$\bar{H}_{n+1} = \bar{H}_n + \frac{1}{n+2} (X_{n+1} X_{n+1}^T - \bar{H}_n)$$

$$\hat{\sigma}_{n+1}^2 = \hat{\sigma}_n^2 + \frac{1}{n+2} \left((Y_{n+1} - X_{n+1}^T \theta_n)^2 - \hat{\sigma}_n^2 \right)$$

SIMULATIONS

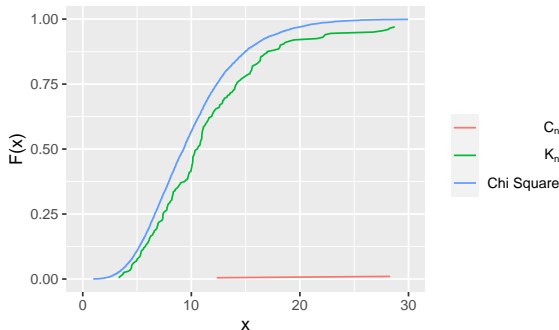


FIGURE – Comparaison des fonctions de répartition de C_n et K_n , pour $n = 5000$, et de la fonction de répartition d'une Chi 2 à 10 degrés de liberté dans le cadre du modèle linéaire.

$$\text{TESTER } x_0^T \theta = x_0^T \theta_0$$

Réécriture du TLC

$$\sqrt{n} \frac{x_0^T \theta_n - x_0^T \theta}{\sqrt{\sigma^2 x_0^T H^{-1} x_0}} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Application : Sous H_0 ,

$$\sqrt{n} \frac{x_0^T \theta_n - x_0^T \theta}{\sqrt{\hat{\sigma}_n^2 x_0^T \bar{H}^{-1} x_0}} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

SIMULATIONS

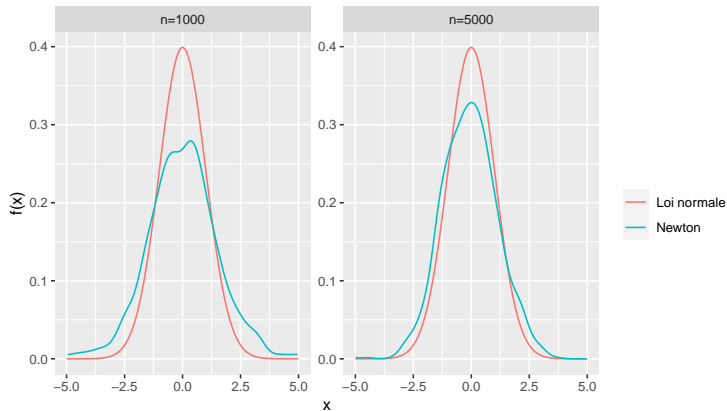


FIGURE – Comparaison de la densité de C_{e1} , pour $n = 1000$ (à gauche) et $n = 5000$ (à droite), et de la densité d'une loi normal centrée réduite dans le cadre de la régression linéaire.

Régression logistique

L'ALGORITHME

Algorithme de Newton stochastique :

$$\alpha_{n+1} = \pi(\theta_n^T X_{n+1}) (1 - \pi(\theta_n^T X_{n+1}))$$

$$\theta_{n+1} = \theta_n + \frac{1}{n+1} \bar{H}_n^{-1} (Y_{n+1} - \pi(\theta_n^T X_{n+1})) X_{n+1}$$

$$H_{n+1}^{-1} = H_n^{-1} - \alpha_{n+1} (1 + \alpha_{n+1} X_{n+1}^T H_n^{-1} X_{n+1})^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}$$

avec H_0^{-1} symétrique et définie positive, $\bar{H}_n^{-1} = (n+1)H_n$.

Réécriture de \bar{H}_n :

$$\bar{H}_n = \frac{1}{n+1} \left(H_0 + \sum_{k=1}^n \pi(\theta_n^T X_{k+1}) (1 - \pi(\theta_n^T X_{k+1})) X_k X_k^T \right)$$

L'ALGORITHME

Algorithme de Newton stochastique tronqué :

$$\alpha_{n+1} = \pi \left(\theta_n^T X_{n+1} \right) \left(1 - \pi \left(\theta_n^T X_{n+1} \right) \right)$$

$$\theta_{n+1} = \theta_n + \frac{1}{n+1} \bar{H}_n^{-1} \left(Y_{n+1} - \pi \left(\theta_n^T X_{n+1} \right) \right) X_{n+1}$$

$$H_{n+1}^{-1} = H_n^{-1} - a_{n+1} \left(1 + a_{n+1} X_{n+1}^T H_n^{-1} X_{n+1} \right)^{-1} H_n^{-1} X_{n+1} X_{n+1}^T H_n^{-1}$$

avec $a_{n+1} = \max \left\{ \alpha_{n+1}, \frac{c_\beta}{(n+1)^\beta} \right\}$ avec $c_\beta > 0$ et $\beta \in (0, 1/2)$

Réécriture de \bar{H}_n :

$$\bar{H}_n = \frac{1}{n+1} \left(H_0 + \sum_{k=1}^n \max \left\{ \alpha_{k+1}, \frac{c_\beta}{(k+1)^\beta} \right\} X_k X_k^T \right)$$

VITESSE DE CONVERGENCE

Théorème

On suppose que X admet un moment d'ordre 4. Alors pour tout $\delta > 0$,

$$\|\theta_n - \theta\|^2 = o\left(\frac{(\ln n)^{1+\delta}}{n}\right) \text{ p.s. } \quad \text{et} \quad \sqrt{n}(\theta_n - \theta) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, H^{-1})$$

SIMULATIONS

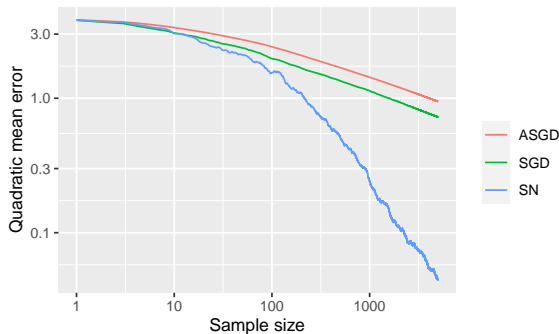


FIGURE – Evolution de l'erreur quadratique moyenne des estimateurs de gradient (SGD), de leur version moyennée (ASGD) et des estimateurs de Newton stochastique (SN) en fonction de la taille de l'échantillon dans le cadre de la régression logistique.

TESTER $H_0 : \theta = \theta_0$ "EN LIGNE"

Réécriture du TLC : Sous H_0 ,

$$\sqrt{n}(\theta_n - \theta_0)^T H(\theta_n - \theta_0) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi_d^2$$

Application : Sous H_0 ,

$$K_n := \sqrt{n}(\theta_n - \theta_0)^T \bar{H}_n(\theta_n - \theta_0) \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \chi_d^2$$

SIMULATIONS

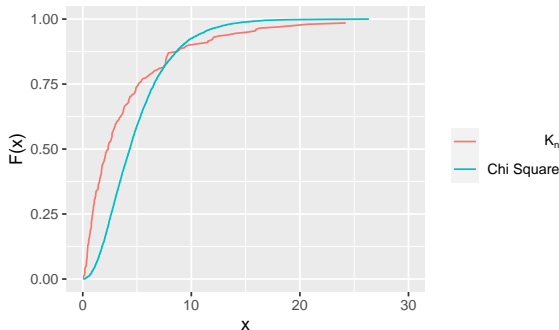


FIGURE – Comparaison de la fonction de répartition de K_n , pour $n = 5000$, et de la fonction de répartition d'une Chi 2 à 10 degrés de liberté.

TESTER $x_0^T \theta = x_0^T \theta_0$

Réécriture du TLC

$$\sqrt{n} \frac{x_0^T \theta_n - x_0^T \theta}{\sqrt{x_0^T H^{-1} x_0}} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

Application : Sous H_0 ,

$$\sqrt{n} \frac{x_0^T \theta_n - x_0^T \theta}{\sqrt{x_0^T \bar{H}^{-1} x_0}} \xrightarrow[n \rightarrow +\infty]{\mathcal{L}} \mathcal{N}(0, 1).$$

SIMULATIONS

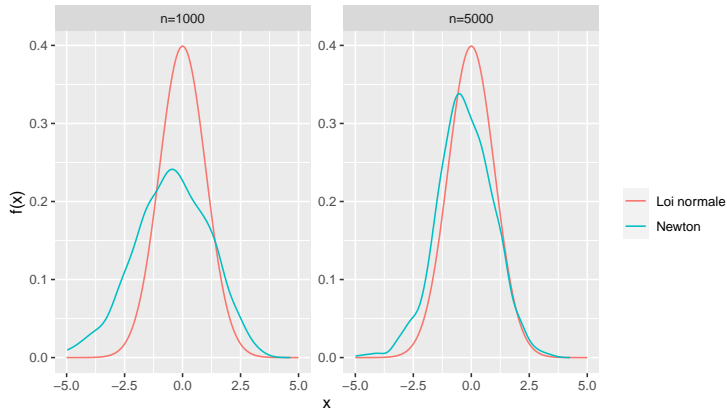


FIGURE – Comparaison de la densité de C_{e_1} , pour $n = 1000$ (à gauche) et $n = 5000$ (à droite), et de la densité d'une loi normale centrée réduite.