Algorithmes de Newton stochastiques

A. Godichon-Baggioni

Algorithme de Newton stochastique

IDÉE

Modèle linéaire :

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

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

$$\begin{aligned}
\mathbb{E}\left[\theta_{n+1} - \theta\right] &= \mathbb{E}\left[\theta_n - \theta\right] - \gamma_{n+1}\mathbb{E}\left[\nabla G\left(\theta_n\right)\right] \\
&= \left(I_d - \gamma_{n+1} H\right)\mathbb{E}\left[\theta_n - \theta\right] \\
&= \begin{pmatrix} 1 - \gamma_{n+1}10^2 & 0 \\ 0 & 1 - \gamma_{n+1}10^{-2} \end{pmatrix}\mathbb{E}\left[\theta_n - \theta\right]
\end{aligned}$$

Algorithme de Newton stochastique

Algorithme de Newton stochastique :

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

Hypothèses sur $\overline{H}_n$:

- $\overline{H}_n^{-1}$ est symétrique et définie positive.
- Il existe une filtration $\left(\mathcal{F}_n\right)$ telle que
  - $\overline{H}_n$ et m_n sont $\mathcal{F}_n = \sigma\left(X_1, \ldots, X_n\right)$ 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

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

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

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

- **(PS5)** $\Longrightarrow \nabla G(.)$ est $L_{\nabla G}$ Lipchitz.
- G fortement convexe + **(PS0)** $\Longrightarrow$ **(PS0")**.
- **(PS0")** + **(PS5)** $\Longrightarrow$ **(PS0)** .

Hypothèse sur l'estimateur $\overline{H}_n$

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

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

avec $\beta < 1/2$.

- **(H1)** $\Longrightarrow \liminf \lambda_{\min}\left(\overline{H}_n^{-1}\right) > 0$ p.s.

Convergence

Théorème

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

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

NOUVELLE HYPOTHÈSE SUR $\overline{H}_n$

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

$$\overline{H}_n \xrightarrow[n\to+\infty]{p.s} H \qquad \text{et} \qquad \overline{H}_n^{-1} \xrightarrow[n\to+\infty]{p.s} H^{-1}$$

- $m_n \xrightarrow[n\to+\infty]{p.s} m \Longrightarrow \overline{H}_n \xrightarrow[n\to+\infty]{p.s} H.$

Vitesse de convergence

Théorème

On suppose que les hypothèses ***(PS0″) (PS2)***, ***(PS4)***, ***(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)**, **(PS4)**, **(PS5)**, **(H1)** et **(H2)** sont vérifiées, alors il existe $p_H > 0$ tel que

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

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

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

Efficacité asymptotique

Théorème

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

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

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

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

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

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

L'algorithme

Algorithme de Newton stochastique :

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

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

Réécriture de $\overline{H}_n$:

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

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 $\delta > 0$,

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

SIMULATIONS

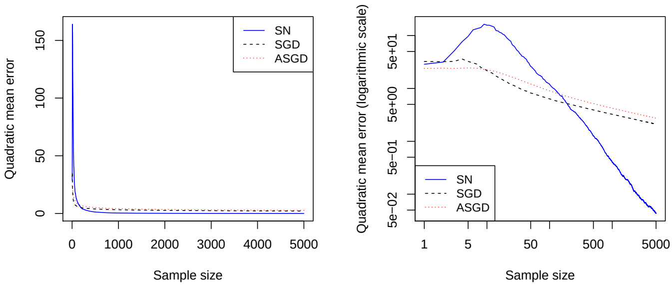


FIGURE – Evolution de l'erreur quadratique moyenne des estimateurs de gradient θ_n (SGD), de leur version moyennée $\overline{\theta}_n$ (ASGD) et des estimateurs de Newton stochastique $\tilde{\theta}_n$ (SN) en fonction de la taille de l'échantilon dans le cadre du modèle linéaire.

TESTER H0 : $\theta = \theta_0$ "EN LIGNE"

Réécriture du TLC : Sous H0,

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

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

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

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

Ecriture directe :

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

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

Ecriture récursive :

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

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

SIMULATIONS

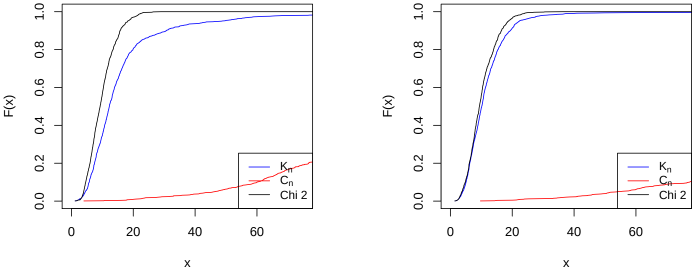


FIGURE – Comparaison des fonctions de répartition de C_n et K_n, pour $n = 1000$ (à gauche) et $n = 5000$ (à droite), et de la fonction de répartition d'une Chi 2 à 10 degrés de liberté dans le cadre du modèle linéaire.

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

Réécriture du TLC

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

Application : Sous H0,

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

SIMULATIONS

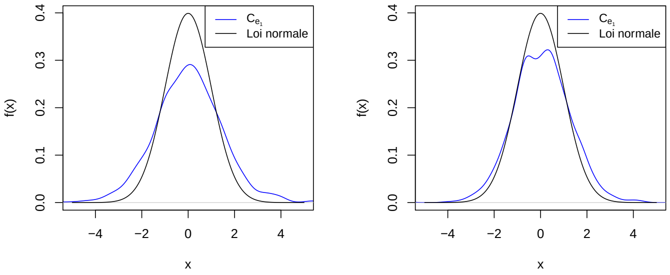


FIGURE – Comparaison de la densité de C_{e_1}, pour $n = 1000$ (à gauche) et $n = 5000$ (à droite), et de la densité d'une loi normal centrée réduite dans le cadre de la régression linéaire.

Régression logistique

L'ALGORITHME

Algorithme de Newton stochastique :

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

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

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

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

Réécriture de $\overline{H}_n$:

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

L'ALGORITHME

Algorithme de Newton stochastique tronqué :

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

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

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

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 $\overline{H}_n$:

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

Vitesse de convergence

Théorème

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

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

SIMULATIONS

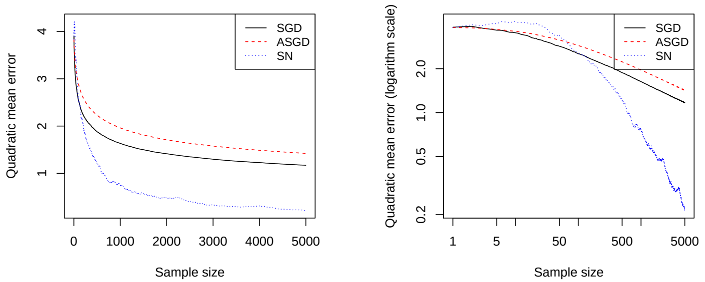


FIGURE – Evolution de l'erreur quadratique moyenne des estimateurs de gradient (SGD), de leur version moyennée (ASGD) et des estimateurs de Newton stochastique (SN) en fonction de la taille de l'échantillon dans le cadre de la régression logistique.

Tester H0 : $\theta = \theta_0$ "en ligne"

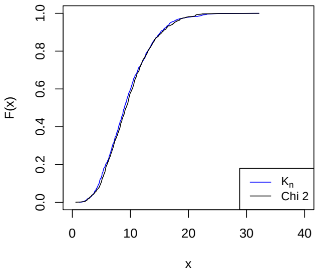
Réécriture du TLC : Sous H0,

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

Application : Soit $\overline{H}_n$ un estimateur consistant de H. Alors

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

SIMULATIONS



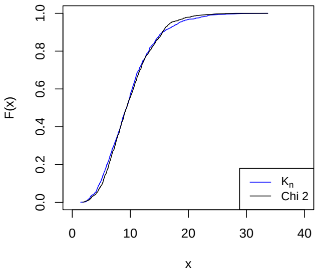


FIGURE – Comparaison de la fonction de répartition de K_n, pour $n = 1000$ (à gauche) et $n = 5000$ (à droite), et de la fonction de répartition d'une Chi 2 à 10 degrés de liberté.

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

Réécriture du TLC

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

Application : Sous H0,

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

SIMULATIONS

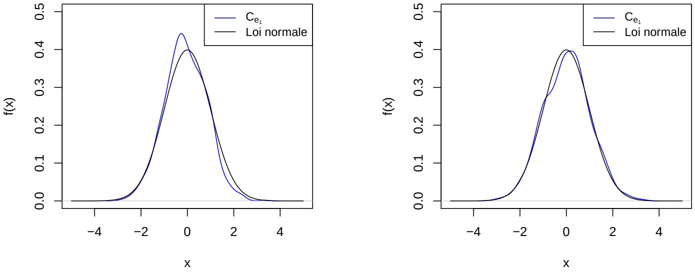


FIGURE – Comparaison de la densité de C_{e_1}, pour $n = 1000$ (à gauche) et $n = 5000$ (à droite), et de la densité d'une loi normal centrée réduite.

Exercice

- On considère le cas de la régression linéaire

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

  avec $\theta = (-2, -1, 0, 1, 2)$, $X \sim \mathcal{N}(0, I_5)$ et $\epsilon \sim \mathcal{N}(0,1)$. Sur un même graphique, tracer l'évolution de l'erreur quadratique moyenne de l'algorithme de gradient, de sa version moyennée et de l'algorithme de Newton stochastique .
- Faire de même mais en prenant $X \sim \mathcal{N}(0, D)$ avec $D = \mathrm{diag}\left(\sigma_i^2\right)$ et $\sigma_i^2 = \frac{i^2}{5^2}$.
- Faire de même pour la régression logistique avec $\theta = (-2, -1, 0, 1, 2)$ et $X \sim (U[0,1])^{\otimes 5}$.