Least squares regression with a random design and Gram matrix estimates

Olivier Catoni

CREST – EXCESS,
CNRS UMR 9194
Olivier.Catoni@ensae.fr

CREST seminar,

Laboratoire de Statistiques

3, avenue Pierre Larousse
92240 Malakoff France

January 25, 2016

Least squares regression

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be n copies of $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$.

$$\begin{aligned}
&\theta_* \in \arg\min_{\theta \in \mathbb{R}^d} \mathbb{E}\big[\big(Y - \langle \theta, X \rangle\big)^2\big],\\
&\widehat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \big(Y_i - \langle \theta, X_i \rangle\big)^2,\\
&G = \mathbb{E}\big(XX^\top\big),\\
&\overline{G} = \frac{1}{n} \sum_{i=1}^n X_i X_i^\top,\\
&R(\theta) = \mathbb{E}\big[\big(Y - \langle \theta, X \rangle\big)^2\big].
\end{aligned}$$

- Bounding $R(\widehat{\theta}) - R(\theta_*)$?
- Replacing $\widehat{\theta}$ by something else ?

Assume that

$$\kappa = \sup\Big\{\mathbb{E}\big[\langle\theta, X\rangle^4\big] \,:\, \theta \in \mathbb{R}^d, \mathbb{E}(\langle\theta, X\rangle^2) \leq 1\Big\} < \infty$$

and that $\quad \mathbb{E}\big[\big(Y - \langle\theta_*, X\rangle\big)^2 \|G^{-1/2}X\|^2\big] < \infty,$

For any $n \geq n_\epsilon = \mathrm{O}_{d\to\infty}\Big(\kappa\big[d + \log(\epsilon^{-1})\big]\Big)$, there is an event Ω of probability at least $1-\epsilon$ such that

$$\begin{aligned}
\mathbb{E}\Big[\big(R(\widehat{\theta}) - R(\theta_*)\big)\mathbb{1}_\Omega\Big] &\leq \frac{(1+\delta)^2}{n}\mathbb{E}\Big[\big(Y - \langle\theta_*, X\rangle\big)^2 \|G^{-1/2}X\|^2\Big] \\
&= \left(1 + \mathrm{O}_{n\to\infty}\left(\sqrt{\frac{\kappa\big[d + \log(\epsilon^{-1})\big]}{n}}\right)\right) \\
&\qquad \times \frac{1}{n}\mathbb{E}\Big[\big(Y - \langle\theta_*, X\rangle\big)^2 \|G^{-1/2}X\|^2\Big].
\end{aligned}$$

Where

$$\delta = \frac{\widehat{\delta} + \gamma_-}{1 - \gamma_-}, \qquad \text{with } \widehat{\delta} = \frac{\mu}{1 - 2\mu},$$

$$\mu = \sqrt{\frac{2(\kappa - 1)}{n}\left[\log(\epsilon^{-1}) + 0.73\, d\right]} + 6.81\sqrt{\frac{2\kappa\, d}{n}}$$

$$\gamma_- = \frac{2\left[0.73\, d + \log(\epsilon^{-1})\right]}{3(\kappa - 1)n}$$

Assume moreover that

$$Y = \langle \theta_*, X \rangle + \eta$$

where η is independent from X, $\mathbb{E}(\eta) = 0$ and $\mathbb{E}\big(\eta^2\big) = \sigma^2 < \infty$. With probabilitiy at least $1 - \epsilon$,

$$\begin{aligned}\mathbb{E}\big[R(\widehat{\theta}) - R(\theta_*)\,\big|\,X_1, \dots, X_n\big] &\leq \frac{(1+\delta)\, d\, \sigma^2}{n} \\ &= \left(1 + \mathrm{O}_{n\to\infty}\left(\sqrt{\frac{\kappa\big[d + \log(\epsilon^{-1})\big]}{n}}\,\right)\right)\frac{d\,\sigma^2}{n}.\end{aligned}$$

When the noise η is Gaussian (but not necessarily the design X), with probability at least $1-2\epsilon$,

$$R(\widehat{\theta}) - R(\theta_*) \leq \frac{(1+\delta)\sigma^2}{n}\big[1.4d + 4\log(\epsilon^{-1})\big].$$

Lower bound

Let $(\widetilde{X}, \widetilde{Y}) \in \mathbb{R}^d \times \mathbb{R}$ be a Gaussian vector such that $\widetilde{Y} = \langle \theta_*, \widetilde{X} \rangle + \eta$, where η is independent from $\widetilde{X}$ and such that $\mathbf{rank}\Big(\mathbb{E}(\widetilde{X}\widetilde{X}^\top)\Big) = d$.

Let σ be an independent Bernoulli random variable of parameter p.

Let $(X, Y) = (\sigma \widetilde{X}, \sigma \widetilde{Y})$.

Remark that $\kappa = 3/p$ and define $\kappa' = \dfrac{\mathbb{E}\big[Y - \langle \theta_*, X \rangle)^4\big]}{\mathbb{E}\big[Y - \langle \theta_*, X \rangle)^2\big]^2} = 3/p$.

For any $n \geq \mathrm{O}_{d \to \infty}\Big((\kappa + \kappa')\big[d + \log(\epsilon^{-1})\big]^3\Big)$, with probability at least $1 - 5\epsilon$,

$$R(\widehat{\theta}) - R(\theta_*) \geq \frac{35}{1000} \times \frac{\big[d - 3\log(\epsilon^{-1})\big](\kappa + \kappa')}{n} R(\theta_*).$$

$$\text{Define } \widetilde{\gamma}_+ = \frac{1}{p+1}\left(\frac{2\big[\log(\epsilon^{-1}) + 0.73(d+1)\big]}{[(\sqrt{\kappa}+\sqrt{\kappa'})^2-1)n}\right)^{p/2}(1+\widehat{\delta})^{p+1}$$

$$\times\Bigg[\left(\frac{\mathbb{E}\big[\big(Y-\langle\theta_*,X\rangle\big)^{2p+2}\big]^{1/(p+1)}}{\mathbb{E}\big[\big(Y-\langle\theta_*,X\rangle\big)^2\big]} + \mathbb{E}\big(\|G^{-1/2}X\|^{2p+2}\big)^{1/(p+1)}\right)^{p+1}$$

$$+\frac{C_q}{\epsilon^{1/q}n^{1-1/q}}\Bigg(\frac{\mathbb{E}\big[\big(Y-\langle\theta_*,X\rangle\big)^{2q(p+1)}\big]^{q^{-1}(p+1)^{-1}}}{\mathbb{E}\big[\big(Y-\langle\theta_*,X\rangle\big)^2\big]}$$

$$+\mathbb{E}\Big(\|G^{-1/2}X\|^{2q(p+1)}\Big)^{q^{-1}(p+1)^{-1}}\Bigg)^{p+1}\Bigg],$$

where $p \in]1,2]$, $q \in]1,2[$ and

$$C_q = \frac{q^{q-1}}{2(q-1)^{q-1}(1-q/2)^{(2-q)/q}}.$$

With probability at least $1-4\epsilon$,

$$R(\widehat{\theta}) - R(\theta_*) \leq \frac{\delta^2}{(1-\delta)^2} R(\theta_*),$$

$$= \mathrm{O}_{n\to\infty}\left(\frac{(\kappa+\kappa')\big[d+\log(\epsilon^{-1})\big]}{n}\right), \text{ where}$$

$$\delta = \frac{\widehat{\delta}+\widetilde{\gamma}_+}{(1-\widehat{\delta})(1-\widetilde{\gamma}_+)}$$

Robust estimator

Consider the bound

$$\mu = \sqrt{\frac{2\big[(\kappa^{1/2}+\kappa'^{1/2})^2-1\big]\big[0.73\,(d+1)+\log(\epsilon^{-1})\big]}{n}} + 6.81\,(\kappa^{1/2}+\kappa'^{1/2})\sqrt{\frac{2(d+1)}{n}}.$$

There is a robust estimator $\widetilde{\theta}$ of θ_* such that for $n \geq \mathrm{O}_{d\to\infty}\Big((\kappa+\kappa')\big[d+\log(\epsilon^{-1})\big]\Big)$, with probability at least $1-2\epsilon$

$$R(\widetilde{\theta}) - R(\theta_*) \leq \frac{\delta^2}{(1-\delta)^2}R(\theta_*)$$

where

$$\delta^2 = \left(\frac{2\mu}{1-4\mu}\right)^2 = \mathrm{O}_{n\to\infty}\left(\frac{(\kappa+\kappa')\big[d+\log(\epsilon^{-1})\big]}{n}\right).$$

Link with Gram matrix estimates

Consider the Gram matrix

$$\widetilde{G} = \mathbb{E}\left[\begin{pmatrix} X \\ -Y \end{pmatrix} (X^\top, -Y)\right],$$

Assume that $\widehat{G}$ is an estimate of $\widetilde{G}$ such that with probability at least $1-\epsilon$, for any $(\theta,\xi) \in \mathbb{R}^d \times \mathbb{R}$,

$$\left| \frac{(\theta^\top,\xi)\,\widehat{G}(\theta^\top,\xi)^\top}{(\theta^\top,\xi)\,\widetilde{G}(\theta^\top,\xi)^\top} - 1 \right| \leq \delta.$$

Consider

$$\widehat{\theta} \in \arg\min_{\theta\in\mathbb{R}^d} (\theta^\top,1)\,\widehat{G}(\theta^\top,1)^\top.$$

With probability at least $1-\epsilon$,

$$\begin{aligned}(\widehat{\theta}^\top,1)\,\widehat{G}(\widehat{\theta}^\top,1)^\top \leq R(\widehat{\theta}) - R(\theta_*) &\leq \frac{\delta^2}{(1-\delta)(1-\delta^2)}(\widehat{\theta}^\top,1)\,\widehat{G}(\widehat{\theta}^\top,1)^\top \\ &\leq \frac{\delta^2}{(1-\delta)^2} R(\theta_*)\end{aligned}$$

Robust Gram matrix estimate

$$\widehat{N}_\lambda(\theta) = \inf\left\{\rho \in \mathbb{R}_+^*, \sum_{i=1}^n \psi\Big[\lambda\Big(\rho^{-1}\langle\theta, X_i\rangle^2 - 1\Big)\Big] \leq 0\right\}$$

$$\lambda = \sqrt{\frac{2}{(\kappa-1)n}\left[0.73d + \log(\epsilon^{-1})\right]},$$

$$n > \left[20\sqrt{\kappa d} + \left(\frac{5}{2} + \frac{1}{2(\kappa-1)}\right)\sqrt{2(\kappa-1)\left[0.73d + \log(\epsilon^{-1})\right]}\right]^2.$$

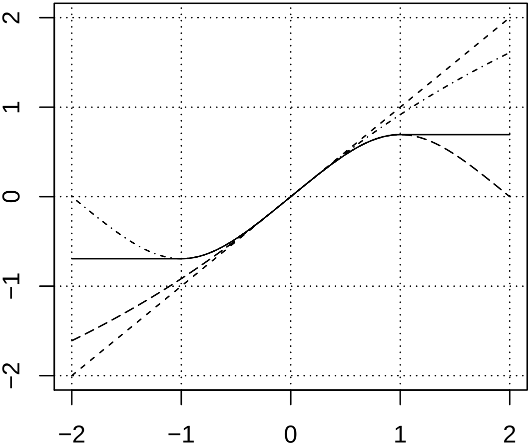
With probability at least $1 - 2\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\left|\frac{\theta^\top G\theta}{\widehat{N}_\lambda(\theta)} - 1\right| \leq \frac{\mu}{1-2\mu} = \mathrm{O}_{n\to\infty}\left(\sqrt{\frac{\kappa\left[d + \log(\epsilon^{-1})\right]}{n}}\right),$$

with the convention that $0/0 = 1$ and $z/0 = \infty$ when $z > 0$.

The influence function ψ is defined as

$$\psi(z) = \begin{cases} \log(2), & z \geq 1, \\ -\log\big(1 - z + z^2/2\big), & 0 \leq z \leq 1, \\ -\psi(-z), & z \leq 0. \end{cases}$$



$z \mapsto \psi(z)$, compared with $z \mapsto z$

$z \mapsto \log(1 + z + z^2/2)$, and $z \mapsto -\log(1 - z + z^2/2)$

Comparison with the empirical Gram matrix

With probabiilty at least $1-\epsilon$,

$$\frac{\theta^\top \overline{G}\theta}{\widehat{N}_\lambda(\theta)} - 1 \leq \inf_{p\in[0,2]} \frac{\lambda^p}{p+1}\frac{1}{n}\sum_{i=1}^n \Bigl(\langle \theta, X_i\rangle \widehat{N}_\lambda(\theta)^{-1} - 1\Bigr)_+^{p+1}$$

$$1 - \frac{\theta^\top \overline{G}\theta}{\widehat{N}_\lambda(\theta)} \leq \frac{\lambda^2}{3} = \gamma_-.$$

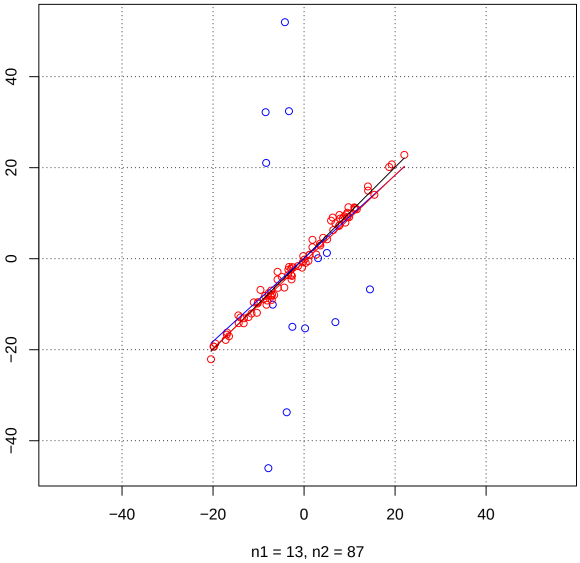
Consequently with probability at least $1-\epsilon$, for any $\theta \in \mathbb{R}^d$,

$$\frac{1-\gamma_-}{1+\widehat{\delta}} \leq \frac{\theta^\top \overline{G}\theta}{\theta^\top G\theta} = \frac{\frac{1}{n}\sum_{i=1}^n \langle \theta, X_i\rangle^2}{\mathbb{E}\bigl(\langle \theta, X\rangle^2\bigr)}.$$

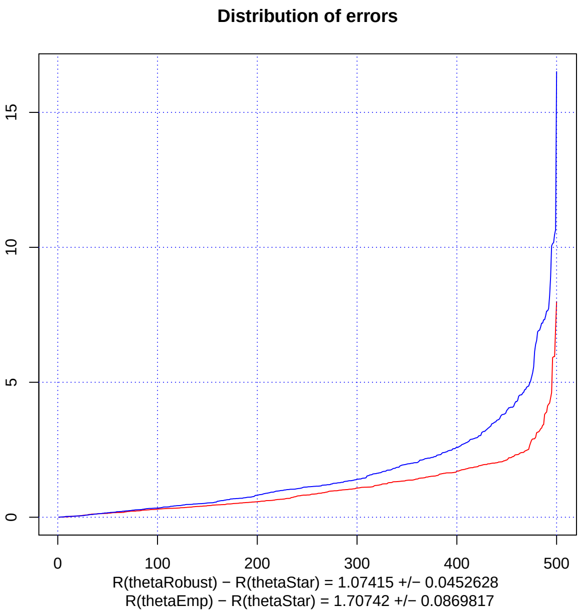
Some simulation

Consider some noise $\eta \sim 0.9 \times \mathscr{N}(0,1) + 0.1 \times \mathscr{N}(0,30^2)$, $X \sim \mathscr{N}(0,10^2)$ independent from η and

$$Y = X + \eta + 1.$$



40
20
0
−20
−40
−40
−20
0
20
40
n1 = 13, n2 = 87


Distribution of errors
0
5
10
15
0
100
200
300
400
500
R(thetaRobust) − R(thetaStar) = 1.07415 +/− 0.0452628
R(thetaEmp) − R(thetaStar) = 1.70742 +/− 0.0869817