Robust regression via Maximum Mean Discrepancy

Pierre Alquier

Center for
Advanced Intelligence Project

New Trends in Statistical Learning II
Porquerolles, June 2022

Contents

Contents

The Maximum Likelihood Estimator (MLE)

Let $X_1, \dots, X_n$ be i.i.d in $\mathcal{X}$ from a probability distribution P_0.

Statistical inference :

- propose a model $(P_\theta, \theta \in \Theta)$, assume $P_0 = P_{\theta_0}$.
- compute $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$.

Letting p_θ denote the density of P_θ, then

$$\hat{\theta}_n^{MLE} = \arg\max_{\theta \in \Theta} L_n(\theta), \text{ where } L_n(\theta) = \prod_{i=1}^{n} p_\theta(X_i).$$

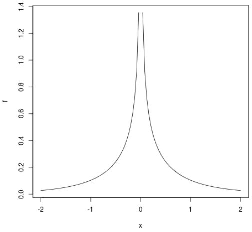
Example : $P_{(m,\sigma)} = \mathcal{N}(m, \sigma^2)$ then

$$\hat{m} = \frac{1}{n} \sum_{i=1}^{n} X_i \text{ and } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{m})^2.$$

MLE not unique / not consistent

Example :

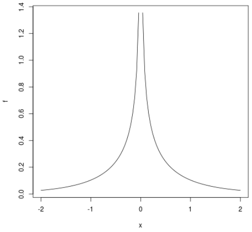
$$p_\theta(x) = \frac{\exp(-|x-\theta|)}{2\sqrt{\pi|x-\theta|}},$$



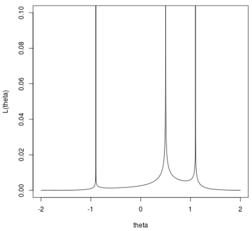
MLE not unique / not consistent

Example :

$$p_\theta(x) = \frac{\exp(-|x-\theta|)}{2\sqrt{\pi|x-\theta|}},$$



$$L_n(\theta) = \frac{\exp\left(-\sum_{i=1}^n |X_i-\theta|\right)}{(2\sqrt{\pi})^n \prod_{i=1}^n \sqrt{|X_i-\theta|}}.$$



MLE fails in the presence of outliers

What is an outlier ?

Huber proposed the contamination model : with probability ε, X_i is not drawn from P_{θ_0} but from Q that can be anything :

$$P_0 = (1-\varepsilon)P_{\theta_0} + \varepsilon Q.$$

Example : $P_\theta = \mathcal{U}nif[0,\theta]$, then

$$L_n(\theta) = \frac{1}{\theta^n}\prod_{i=1}^{n} \mathbb{1}_{\{0\leq X_i \leq \theta\}} \Rightarrow \hat{\theta} = \max_{1\leq i\leq n} X_i.$$

In the case of the following contamination, the MLE is extremely far from the truth :

$$P_0 = (1-\varepsilon).\mathcal{U}nif[0,1] + \varepsilon.\mathcal{N}(10^{10},1)...$$

Minimum Distance Estimation

$$\text{Empirical distribution : } \hat{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

Minimum Distance Estimation (MDE)

Let $d(\cdot,\cdot)$ be a metric on probability distributions.

$$\hat{\theta}_d := \arg\min_{\theta\in\Theta} d\left(P_\theta, \hat{P}_n\right).$$

Wolfowitz, J. (1957). The minimum distance method. *The Annals of Mathematical Statistics*.

Idea : MDE with an adequate d leads to robust estimation.

Bickel, P. J. (1976). Another look at robustness : a review of reviews and some new developments. *Scandinavian Journal of Statistics*. Discussion by Sture Holm.

Parr, W. C. & Schucany, W. R. (1980). Minimum distance and robust estimation. *JASA*.

Yatracos, Y. G. (1985). Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Annals of Statistics*.

Integral Probability Semimetrics

Integral Probability Semimetrics (IPS)

Let $\mathcal{F}$ be a set of real-valued, measurable functions and put

$$d_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{X\sim Q}[f(X)] \right|.$$

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Applied Probability*.

- assumptions required in order to ensure that $d_{\mathcal{F}}(P, Q) = 0 \Rightarrow P = Q$ (that is, $d_{\mathcal{F}}$ is a metric).
- assumptions required in order to ensure that $d_{\mathcal{F}} < +\infty$.

Non-asymptotic bound for MDE

Theorem 1

- $X_1, \dots, X_n$ i.i.d from P_0,
- for any $f \in \mathcal{F}$, $\sup_{x\in\mathcal{X}} |f(x)| \leq 1$.

Then

$$\mathbb{E}\left[d_{\mathcal{F}}(P_{\hat{\theta}_{d_{\mathcal{F}}}}, P_0)\right] \leq \inf_{\theta\in\Theta} d_{\mathcal{F}}(P_\theta, P_0) + 4.\mathrm{Rad}_n(\mathcal{F}).$$

Rademacher complexity

$$\mathrm{Rad}_n(\mathcal{F}) := \sup_P \mathbb{E}_{Y_1,\dots,Y_n\sim P}\, \mathbb{E}_{\epsilon_1,\dots,\epsilon_n}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(Y_i)\right].$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d Rademacher variables :

$$\mathbb{P}(\epsilon_1 = 1) = \mathbb{P}(\epsilon_1 = -1) = 1/2.$$

Example 1 : set of indicators

$$\mathbb{1}_A(x) = \begin{cases} 1 \text{ if } x \in A, \\ 0 \text{ if } x \notin A. \end{cases}$$

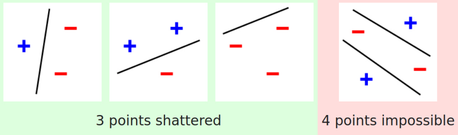


Image from Wikipedia.

Reminder - Vapnik-Chervonenkis dimension

Assume that $\mathcal{F} = \{\mathbb{1}_A, A \in \mathcal{A}\}$ for some $\mathcal{A} \subseteq \mathcal{P}(\mathcal{X})$,

- $S_{\mathcal{F}}(x_1, \dots, x_n) := \{(f(x_1), \dots, f(x_n)), f \in \mathcal{F}\}$,
- $\mathrm{VC}(\mathcal{F}) := \max\left\{n : \exists x_1, \dots, x_n, \left|S_{\mathcal{F}}(x_1, \dots, x_n)\right| = 2^n\right\}$.

Theorem (Bartlett and Mendelson)

$$\mathrm{Rad}_n(\mathcal{F}) \leq \sqrt{\frac{2.\mathrm{VC}(\mathcal{F}) \log(n+1)}{n}}.$$

Bartlett, P. L. & Mendelson, S. (2002). Rademacher and Gaussian complexities : Risk bounds and structural results. *JMLR*.

Example 1 : KS and TV distances

Two classical examples :

- $\mathcal{A} = \{$all measurable sets in $\mathcal{X}\}$, then $d_{\mathcal{F}}(\cdot,\cdot)$ is the total variation distance $\mathrm{TV}(\cdot,\cdot)$.
  - $\mathrm{VC}(\mathcal{F}) = +\infty$ when $|\mathcal{X}| = +\infty$,
  - in general, $\mathrm{Rad}_n(\mathcal{F}) \nrightarrow 0$.
- $\mathcal{X} = \mathbb{R}$, $\mathcal{A} = \{(-\infty, x], x \in \mathbb{R}\}$, then $d_{\mathcal{F}}(\cdot,\cdot)$ is the Kolmogorov-Smirnov distance $\mathrm{KS}(\cdot,\cdot)$.
  - KS distance was actually proposed by S. Holm for robust estimation,
  - $\mathrm{VC}(\mathcal{F}) = 1$.

$$\mathbb{E}\left[\mathrm{KS}(P_{\hat{\theta}_{\mathrm{KS}}}, P_0)\right] \leq \inf_{\theta\in\Theta} \mathrm{KS}(P_\theta, P_0) + 4.\sqrt{\frac{2\log(n+1)}{n}}.$$

Example 2 : Maximum Mean Discrepancy (MMD)

- Let $(\mathcal{H}, \langle\cdot,\cdot\rangle_{\mathcal{H}})$ be a RKHS with kernel
$$k(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}} .$$
- If $\|\phi(x)\|_{\mathcal{H}} = k(x,x) \leq 1$ then $\mathbb{E}_{X\sim P}[\phi(X)]$ is well-defined .
- The map $P \mapsto \mathbb{E}_{X\sim P}[\phi(X)]$ is one-to-one if k is *characteristic*.
- For example, $k(x,y) = \exp(-\|x-y\|^2/\gamma^2)$ works.

Definition - MMD

$$\begin{aligned}\mathrm{MMD}_k(P,Q) &= \sup_{\substack{f \in \mathcal{H} \\ \|f\|_{\mathcal{H}} \leq 1}} \left| \mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{X\sim Q}[f(X)] \right| \\ &= \left\| \mathbb{E}_{X\sim P}[\phi(X)] - \mathbb{E}_{X\sim Q}[\phi(X)] \right\|_{\mathcal{H}} .\end{aligned}$$

Example 2 : MMD

$$\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\} \Rightarrow \mathrm{Rad}_n(\mathcal{F}) \leq \sqrt{\frac{\sup_x k(x,x)}{n}}.$$

Theorem 2

For k bounded by 1 and characteristic,

$$\mathbb{E}\left[\mathrm{MMD}_k(P_{\hat{\theta}_{\mathrm{MMD}_k}}, P_0)\right] \leq \inf_{\theta \in \Theta} \mathrm{MMD}_k(P_\theta, P_0) + \frac{2}{\sqrt{n}}.$$

Joint work with Badr-Eddine Chérief-Abdellatif (Oxford).

Chérief-Abdellatif, B.-E. and Alquier, P. Finite Sample Properties of Parametric MMD Estimation : Robustness to Misspecification and Dependence. **Bernoulli**, 2022.

Example 2 : MMD

We actually have

$$\mathrm{MMD}_k^2(P_\theta, \hat{P}_n) = \mathbb{E}_{X,X'\sim P_\theta}[k(X,X')] - \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{X\sim P_\theta}[k(X_i,X)] + \frac{1}{n^2}\sum_{1\leq i,j\leq n} k(X_i,X_j)$$

and so

$$\nabla_\theta \mathrm{MMD}_k^2(P_\theta, \hat{P}_n) = 2\mathbb{E}_{X,X'\sim P_\theta}\left\{\left[k(X,X') - \frac{1}{n}\sum_{i=1}^{n} k(X_i,X)\right]\nabla_\theta[\log p_\theta(X)]\right\}$$

that can be approximated by sampling from P_θ.

Example 2 : MMD

Dziugaite, G. K., Roy, D. M., & Ghahramani, Z. (2015). Training generative neural networks via maximum mean discrepancy optimization. *UAI 2015*.

define the estimator and used it to train GANs.

Briol, F. X., Barp, A., Duncan, A. B., & Girolami, M. (2019). Statistical Inference for Generative Models with Maximum Mean Discrepancy. *Preprint arXiv :1906.05944*.

$$\text{assumptions} \Rightarrow \quad \sqrt{n}(\hat{\theta}_{\mathrm{MMD}_k} - \theta_0) \rightsquigarrow \mathcal{N}(0, V_0(k)).$$

Example 3 : Wasserstein

Another classical metric belongs to the IPS family :

$$\mathrm{W}_{\delta}(P,Q) = \sup_{\substack{f:\mathcal{X}\to\mathbb{R} \\ \mathrm{Lip}(f)\leq 1}} \left|\mathbb{E}_{X\sim P}[f(X)] - \mathbb{E}_{X\sim Q}[f(X)]\right|$$

where $\mathrm{Lip}(f) := \sup_{x\neq y} |f(x)-f(y)|/\delta(x,y)$.

- In general, $\mathrm{Rad}_n(\mathcal{F}) \nrightarrow 0$, so will not converge in full generality as with MMD and KS.
- However, nice results can be proven under additional assumptions :

Bernton, E., Jacob, P. E., Gerber, M. & Robert, C. P. (2019). On parameter estimation with the Wasserstein distance. *Information and Inference : A Journal of the IMA*.

MDE and robustness

Reminder

$$\mathbb{E}\left[d_{\mathcal{F}}(P_{\hat{\theta}_{d_{\mathcal{F}}}}, P_0)\right] \leq \inf_{\theta\in\Theta} d_{\mathcal{F}}(P_\theta, P_0) + 4.\mathrm{Rad}_n(\mathcal{F}).$$

Huber's contamination model : $P_0 = (1-\varepsilon)P_{\theta_0} + \varepsilon Q$.

$$\begin{aligned} d_{\mathcal{F}}(P_{\theta_0}, P_0) &= \sup_{f\in\mathcal{F}} \left|\mathbb{E}_{X\sim P_{\theta_0}} f(X) - (1-\varepsilon)\mathbb{E}_{X\sim P_{\theta_0}} f(X) - \varepsilon\mathbb{E}_{X\sim Q} f(X)\right| \\ &= \sup_{f\in\mathcal{F}} \left|\varepsilon\mathbb{E}_{X\sim P_{\theta_0}} f(X) - \varepsilon\mathbb{E}_{X\sim Q} f(X)\right| \\ &= \varepsilon.d_{\mathcal{F}}(P_{\theta_0}, Q) \leq 2\varepsilon \quad \text{if for any } f\in\mathcal{F}, \sup_x |f(x)| \leq 1 \end{aligned}$$

Corollary - in Huber's contamination model

$$\mathbb{E}\left[d_{\mathcal{F}}(P_{\hat{\theta}_{d_{\mathcal{F}}}}, P_{\theta_0})\right] \leq 4\varepsilon + 4.\mathrm{Rad}_n(\mathcal{F}).$$

MDE and robustness : toy experiment

Model : $\mathcal{N}(\theta, 1)$, $X_1, \dots, X_n$ i.i.d $\mathcal{N}(\theta_0, 1)$, $n = 100$ and we repeat the exp. 200 times. Kernel $k(x, y) = \exp(-|x - y|)$.

	$\hat{\theta}_{MLE}$	$\hat{\theta}_{\mathrm{MMD}_k}$	$\hat{\theta}_{\mathrm{KS}}$
mean abs. error	0.081	0.094	0.088

Now, $\varepsilon = 2\%$ of the observations drawn from a Cauchy.

mean abs. error	0.276	0.095	0.088

Now, $\varepsilon = 1\%$ are replaced by $1,000$.

mean abs. error	10.008	0.088	0.082

Contents

Regression with MMD

What follows is based on a joint paper with Mathieu Gerber (Bristol).

Alquier, P. and Gerber, M. (2020). *Universal Robust Regression via Maximum Mean Discrepancy*. Preprint arXiv, submitted.

Examples

We observe independent pairs $(X_1, Y_1), \dots, (X_n, Y_n)$.

1. Gaussian linear regression model :
   - $Y_i = \theta^T X_i + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.
   - can we rewritten as $Y_i \sim \mathcal{N}(\theta^T X_i, \sigma^2)$.
2. Gaussian nonlinear regression model (heteroskedastic) :
   - $Y_i = G(\theta_1^T X_i) + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, (\theta_2^T X_t)^2)$.
   - can we rewritten as $Y_i \sim \mathcal{N}(G(\theta^T X_i), (\theta_2^T X_t)^2)$.
3. Poisson regression : $Y_i = \mathcal{P}(\exp(\theta^T X_i))$.
4. logistic, gamma, binomial, whateveryouwant regression...

Problem in these examples

Regression model

In general, we will consider a parametric model $(P_\theta)_{\theta\in\Theta}$, a function $g : (\Lambda \times \mathcal{X}) \to \Theta$ and

$$Y_i | X_i = x \sim P_{g(\lambda,x)}.$$

Until now, MMD estimation allows us to model the distributions, not conditional distributions.

	truth	model	empirical
X	P^0_X	Q_μ	$\hat{P}^0_X = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$
$Y\|X$	$P^0_{Y\|X}$	$P_{g(\lambda,X)}$	NA
(X,Y)	$P^0 = P^0_X P^0_{Y\|X}$	$R_{(\mu,\lambda)} = Q_\mu P_{g(\lambda,X)}$	$\hat{P}^0 = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_i)}$

Various approaches

	truth	model	empirical
X	P_X^0	Q_μ	$\hat{P}_X^0 = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$
$Y\|X$	$P_{Y\|X}^0$	$P_{g(\lambda,X)}$	NA
(X,Y)	$P^0 = P_X^0 P_{Y\|X}^0$	$R_{(\mu,\lambda)} = Q_\mu P_{g(\lambda,X)}$	$\hat{P}^0 = \frac{1}{n}\sum_{i=1}^n \delta_{(X_i,Y_i)}$

Fix a kernel $k((x,y),(x',y'))$. 3 approaches :

1. completely parametric : $\widehat{(\mu,\lambda)} = \arg\min_{(\mu,\lambda)} \mathrm{MMD}_k(Q_\mu P_{g(\lambda,X)}, \hat{P}^0)$.

   $\rightarrow$ what is the effect of a misspecification in Q_μ?

2. semi-parametric 1 : $\widehat{(Q,\lambda)} = \arg\min_{(Q,\lambda)} \mathrm{MMD}_k(QP_{g(\lambda,X)}, \hat{P}^0)$.

   $\rightarrow$ we dont know how to minimize this.

3. semi-parametric 2 : $\widehat{\lambda} = \arg\min_{\lambda} \mathrm{MMD}_k(\hat{P}_X^0 P_{g(\lambda,X)}, \hat{P}^0)$.

   $\rightarrow$ our approach.

Explicit minimization problem

$$\widehat{\lambda} = \arg\min_{\lambda} C(\lambda) := \mathrm{MMD}_k(\hat{P}^0_X P_{g(\lambda,X)}, \hat{P}^0).$$

$$\begin{aligned} C(\lambda) = \frac{1}{n} \sum_{i,j=1}^{n} & \mathbb{E}_{Y \sim P_{g(\lambda,X_i)}, Y' \sim P_{g(\lambda,X_j)}} k((X_i, Y), (X_j, Y')) \\ & - \frac{2}{n} \sum_{i,j=1}^{n} \mathbb{E}_{Y \sim P_{g(\lambda,X_i)}} k((X_i, Y), (X_j, Y_j)) \\ & + \frac{1}{n} \sum_{i,j=1}^{n} k((X_i, Y_i), (X_j, Y_j)) \end{aligned}$$

We also proposed an approximation $\tilde{\lambda}$ of this estimator by removing the non-diagonal terms. We have an asymptotic theory for $\tilde{\lambda}$.

Theoretical results : fixed design

$$\widehat{\lambda} = \arg\min_{\lambda} C(\lambda) := \mathrm{MMD}_k(\hat{P}^0_X P_{g(\lambda,X)}, \hat{P}^0).$$

Theorem

Assume $|k| \leq 1$. Then

$$\mathbb{E}_{Y_1,\dots,Y_n}\left[\mathrm{MMD}_k(\hat{P}^0_X P_{g(\hat{\lambda},X)}, \hat{P}^0_X P^0_{Y|X})\right] \leq \min_{\lambda} \mathrm{MMD}_k(\hat{P}^0_X P_{g(\lambda,X)}, \hat{P}^0_X P^0_{Y|X}) + \frac{2}{\sqrt{n}}.$$

The result is a relatively easy adaptation of the general theory of MMD.

Theoretical results : random design

Theorem

Assume :

1. $|k| \leq 1$,
2. $k((x,y),(x',y')) = k_1(x,x')k_2(y,y')$, with associated RKHS $\mathcal{H}_1$ and $\mathcal{H}_2$.
3. for any $f_2 \in \mathcal{H}_2$, for any $\lambda \in \Lambda$, $f_1(x) := \mathbb{E}_{Y \sim P_{g(\lambda,x)}}[f_2(Y)]$ satisfies $f_1 \in \mathcal{H}_1$.

There is a constant $C(k_1)$ that depends only on k_1 such that

$$\mathbb{E}\left[\mathrm{MMD}_k(P_X^0 P_{g(\hat{\lambda},X)}, P^0)\right] \leq \min_{\lambda} \mathrm{MMD}_k(P_X^0 P_{g(\lambda,X)}, P^0) + \frac{C(k_1)\sqrt{2}+3}{\sqrt{n}}.$$

Kernels satisfying our conditions

For any $f_2 \in \mathcal{H}_2$, for any $\lambda \in \Lambda$, $f_1(x) := \mathbb{E}_{Y \sim P_{g(\lambda,x)}}[f_2(Y)]$ satisfies $f_1 \in \mathcal{H}_1$.

- not easy.
- famous paper claiming that it is true regardless of $P_{g(\lambda,x)}$, + erratum.
- derivations in our paper to provide examples : 10 pages. Take home message : Matérn kernels + a transformation of the inputs work for all the classical regression models.

Numerical examples : linear regression

τ	type	n	$\beta_{\mathrm{ols},n}$	$\beta_{\mathrm{lad},n}$	$\beta_{\mathrm{rob},n}$	$\hat{\beta}_n$	$\tilde{\beta}_n$
0		100	0.372	0.353	0.350	0.355	0.334
		1 000	0.116	0.092	0.104	0.108	0.107
		5 000	0.053	0.039	0.046	0.049	0.047
1	Y	100	0.464	0.339	0.385	0.350	0.342
		1 000	0.181	0.094	0.106	0.105	0.097
		5 000	0.103	0.043	0.049	0.054	0.051
2	Y	100	0.647	0.351	0.359	0.337	0.333
		1 000	0.241	0.097	0.110	0.114	0.115
		5 000	0.175	0.039	0.047	0.051	0.052
3	Y	100	0.724	0.331	0.343	0.329	0.320
		1 000	0.309	0.100	0.108	0.113	0.110
		5 000	0.250	0.043	0.048	0.053	0.055
1	X	100	0.870	0.356	0.374	0.342	0.338
		1 000	0.836	0.111	0.105	0.104	0.096
		5 000	0.818	0.065	0.049	0.054	0.052
2	X	100	1.575	0.400	0.347	0.337	0.331
		1 000	1.467	0.160	0.110	0.112	0.115
		5 000	1.401	0.119	0.046	0.051	0.052
3	X	100	1.838	0.442	0.344	0.331	0.323
		1 000	1.805	0.216	0.108	0.113	0.109
		5 000	1.771	0.183	0.048	0.054	0.056

Table 1: Results for the Gaussian linear regression model. For each experimental setting we report the RMSE over 25 replications.

Numerical examples : gamma regression

ϵ	n	θ_{mle}	θ_{rob}	$\hat{\theta}_n$	$\tilde{\theta}_n$
0%	100	0.478	0.425	0.501	0.498
	1 000	0.123	0.120	0.142	0.144
	5 000	0.055	0.053	0.067	0.069
1%	100	0.394	0.398	0.521	0.521
	1 000	0.250	0.108	0.149	0.148
	5 000	0.350	0.049	0.071	0.070
2%	100	0.395	0.381	0.539	0.545
	1 000	0.353	0.113	0.146	0.145
	5 000	0.551	0.056	0.073	0.073
3%	100	0.441	0.407	0.471	0.475
	1 000	0.349	0.119	0.157	0.158
	5 000	0.631	0.069	0.076	0.078

ϵ	n	β_{mle}	β_{rob}	$\hat{\beta}_n$	$\tilde{\beta}_n$
0%	100	0.308	0.306	0.444	0.443
	1 000	0.087	0.093	0.132	0.133
	5 000	0.041	0.043	0.061	0.063
1%	100	0.297	0.313	0.473	0.475
	1 000	0.114	0.098	0.132	0.133
	5 000	0.070	0.042	0.061	0.059
2%	100	0.302	0.301	0.490	0.493
	1 000	0.120	0.094	0.127	0.127
	5 000	0.105	0.046	0.066	0.067
3%	100	0.296	0.296	0.403	0.406
	1 000	0.119	0.091	0.138	0.136
	5 000	0.129	0.044	0.067	0.069

Figure 1: Results for the Gamma regression model. For each experimental setting, we report the mean square error over 25 replications.

La fin

終わり

ありがとう ございます。