Parametric estimation via MMD optimization: robustness to outliers and to dependence

Pierre Alquier

Center for
Advanced Intelligence Project

DataSig Seminar
November 2, 2020

RIKEN AIP : ABI team

RIKEN
AIP
Center for
Advanced Intelligence Project

Approximate Bayesian Inference team (ABI), lead by Emtiyaz Khan

Please visit the team website

https ://team-approx-bayes.github.io/

Co-authors

Badr-Eddine Chérief-Abdellatif

Univ. of Oxford

Alexis Derumigny

Univ. of Twente

Mathieu Gerber

Univ. of Bristol

Jean-David Fermanian

ENSAE Paris

The Maximum Likelihood Estimator (MLE)

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

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 $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$.

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

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

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

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

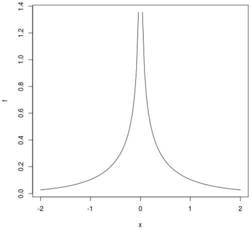
Example : $P_{(m,\sigma)} = \mathcal{N}(m, \sigma^2)$ then

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

MLE not unique / not consistent

Example :

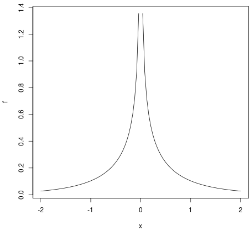
$$p_\theta(x) = \frac{\exp(-|x-\theta|)}{2\sqrt{\pi|x-\theta|}},$$



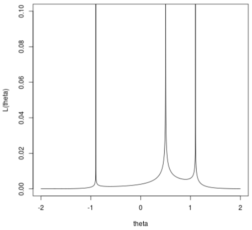
MLE not unique / not consistent

Example :

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



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



MLE fails in the presence of outliers

What is an outlier ?

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 :

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

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

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

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

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

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

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

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

Requirements for a “good” estimator

A universal estimator $\hat{\theta}_n$ must be such that, for some distance d on probability distributions,

Requirements for a "good" estimator

A universal estimator $\hat{\theta}_n$ must be such that, for some distance d on probability distributions,

1. when the model is well specified, that is, $P_0 = P_{\theta_0}$,

$$\mathbb{E}\left[d(P_{\hat{\theta}_n}, P_0)\right] \leq r_n(\Theta) \xrightarrow[n\to\infty]{} 0.$$

Requirements for a "good" estimator

A universal estimator $\hat{\theta}_n$ must be such that, for some distance d on probability distributions,

1. when the model is well specified, that is, $P_0 = P_{\theta_0}$,

$$\mathbb{E}\left[d(P_{\hat{\theta}_n}, P_0)\right] \leq r_n(\Theta) \xrightarrow[n\to\infty]{} 0.$$

2. in the misspecified case $P_0 = (1-\varepsilon)P_{\theta_0} + \varepsilon Q$, for any Q,

$$\mathbb{E}\left[d(P_{\hat{\theta}_n}, P_0)\right] \leq \underbrace{d(P_0, P_{\theta_0})}_{\xrightarrow[\varepsilon\to\infty]{} 0} + \underbrace{r_n(\Theta)}_{\xrightarrow[n\to\infty]{} 0} .$$

Requirements for a “good” estimator

A universal estimator $\hat{\theta}_n$ must be such that, for some distance d on probability distributions,

1. when the model is well specified, that is, $P_0 = P_{\theta_0}$,

$$\mathbb{E}\left[d(P_{\hat{\theta}_n}, P_0)\right] \leq r_n(\Theta) \xrightarrow[n\to\infty]{} 0.$$

2. in the misspecified case $P_0 = (1-\varepsilon)P_{\theta_0} + \varepsilon Q$, for any Q,

$$\mathbb{E}\left[d(P_{\hat{\theta}_n}, P_0)\right] \leq \underbrace{d(P_0, P_{\theta_0})}_{\xrightarrow[\varepsilon\to\infty]{} 0} + \underbrace{r_n(\Theta)}_{\xrightarrow[n\to\infty]{} 0}.$$

The MLE does not satisfy these requirements.

Some examples

Yatracos' skeleton estimate $\hat{\theta}_n^Y$:

$$\mathbb{E}\left[d_{TV}(P_{\hat{\theta}_n^Y}, P_0)\right] \leq 3 d_{TV}(P_0, P_{\theta_0}) + C.\sqrt{\frac{\dim(\Theta)}{n}}$$

where

$$d_{TV}(P, Q) = \sup_E |P(E) - Q(E)|.$$

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

Some examples

Yatracos' skeleton estimate $\hat{\theta}_n^Y$:

$$\mathbb{E}\left[d_{TV}(P_{\hat{\theta}_n^Y}, P_0)\right] \leq 3 d_{TV}(P_0, P_{\theta_0}) + C.\sqrt{\frac{\dim(\Theta)}{n}}$$

where

$$d_{TV}(P, Q) = \sup_E |P(E) - Q(E)|.$$

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

More recent work with the Hellinger distance :

Baraud, Y., Birgé, L., & Sart, M. (2017). A new method for estimation and model selection : ρ-estimation. *Inventiones mathematicae*.

But...

But...

Problem with the aforementioned estimators : they cannot be computed in practice.

But...

Problem with the aforementioned estimators : they cannot be computed in practice.

Additional requirement : an estimator must be computable ! ! !

Overview of the talk

1. Estimation via MMD optimization
   - Definition of the estimator
   - Basic properties
   - References and further works
2. Robustness to outliers
   - Application to Huber contamination model
   - Example : estimation of the mean of a Gaussian
   - Numerical experiments
3. Robustness to dependence
   - Extension to non-independent observations
   - A (new ?) dependence coefficient
   - Example : auto-regressive observations

Reminder : kernels

Let $\mathcal{H}$ be a Hilbert space and any continuous function $\Phi : \mathcal{X} \to \mathcal{H}$. The function

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$$

is called a kernel.

Reminder : kernels

Let $\mathcal{H}$ be a Hilbert space and any continuous function $\Phi : \mathcal{X} \to \mathcal{H}$. The function

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$$

is called a kernel. Conversely :

Mercer's theorem

Let $K(x, y)$ be a continuous function such that for any $(x_1, \ldots, x_n) \in \mathcal{X}^n$ and $(c_1, \ldots, c_n) \neq (0, \ldots, 0) \in \mathbb{R}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j K(x_i, x_j) > 0,$$

then there is $\mathcal{H}$ and Φ such that $K(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}}$.

Reminder : MMD

Assume that the kernel is bounded : $0 \leq K(x,y) \leq 1$.

Reminder : MMD

Assume that the kernel is bounded : $0 \leq K(x, y) \leq 1$.

Consider, for any probability distribution P on $\mathcal{X}$,

$$\mu_K(P) = \mathbb{E}_{X\sim P}\left[\Phi(x)\right].$$

Reminder : MMD

Assume that the kernel is bounded : $0 \leq K(x, y) \leq 1$.

Consider, for any probability distribution P on $\mathcal{X}$,

$$\mu_K(P) = \mathbb{E}_{X \sim P} [\Phi(x)] .$$

The kernel K is said to be characteristic if

$$\mu_K(P) = \mu_K(Q) \Rightarrow P = Q.$$

Reminder : MMD

Assume that the kernel is bounded : $0 \leq K(x,y) \leq 1$.

Consider, for any probability distribution P on $\mathcal{X}$,

$$\mu_K(P) = \mathbb{E}_{X\sim P}\left[\Phi(x)\right].$$

The kernel K is said to be characteristic if

$$\mu_K(P) = \mu_K(Q) \Rightarrow P = Q.$$

Theorem

$K(x,y) = \exp(-\frac{\|x-y\|^2}{\gamma^2})$ and $\exp(-\frac{\|x-y\|}{\gamma})$ are char. kernels.

Reminder : MMD

Assume that the kernel is bounded : $0 \leq K(x,y) \leq 1$.

Consider, for any probability distribution P on $\mathcal{X}$,

$$\mu_K(P) = \mathbb{E}_{X\sim P}\left[\Phi(x)\right].$$

The kernel K is said to be characteristic if

$$\mu_K(P) = \mu_K(Q) \Rightarrow P = Q.$$

Theorem

$K(x,y) = \exp(-\frac{\|x-y\|^2}{\gamma^2})$ and $\exp(-\frac{\|x-y\|}{\gamma})$ are char. kernels.

Definition : the MMD distance

$$\mathbb{D}_K(P,Q) = \left\| \mu_K(P) - \mu_K(Q) \right\|_{\mathcal{H}}.$$

MMD-based estimator

Reminder of the context :

1. $X_1, \dots, X_n$ be i.i.d in $\mathcal{X}$ from a probability distribution P_0,
2. model $(P_\theta, \theta \in \Theta)$.

MMD-based estimator

Reminder of the context :

1. $X_1, \dots, X_n$ be i.i.d in $\mathcal{X}$ from a probability distribution P_0,
2. model $(P_\theta, \theta \in \Theta)$.

Definition - MMD based estimator

$$\hat{\theta}_n^{MMD} = \arg\min_{\theta \in \Theta} \mathbb{D}_K\left(P_\theta, \hat{P}_n\right) \text{ where } \hat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

A preliminary lemma

Lemma

For any P_0, when $X_1, \dots, X_n$ are i.i.d from P_0,

$$\mathbb{E}\left[\mathbb{D}_K\left(\hat{P}_n, P^0\right)\right] \leq \frac{1}{\sqrt{n}}.$$

A preliminary lemma

Lemma

For any P_0, when $X_1, \dots, X_n$ are i.i.d from P_0,

$$\mathbb{E}\left[\mathbb{D}_K\left(\hat{P}_n, P^0\right)\right] \leq \frac{1}{\sqrt{n}}.$$

$$\begin{aligned}\left\{\mathbb{E}\left[\mathbb{D}_K\left(\hat{P}_n, P^0\right)\right]\right\}^2 &\leq \mathbb{E}\left[\mathbb{D}_K^2\left(\hat{P}_n, P^0\right)\right] \\ &= \mathbb{E}\left[\left\|(1/n)\sum(\mu(\delta_{X_i}) - \mu(P_0))\right\|_{\mathcal{H}}^2\right] \\ &= (1/n)\mathbb{E}\left[\|\mu(\delta_{X_1}) - \mu(P_0)\|_{\mathcal{H}}^2\right] \\ &\leq 1/n.\end{aligned}$$

A bound in expectation

$$
\begin{aligned}
\forall\theta,\ \mathbb{D}_K\left(P_{\hat{\theta}_n^{MMD}}, P^0\right) &\leq \mathbb{D}_K\left(P_{\hat{\theta}_n^{MMD}}, \hat{P}_n\right) + \mathbb{D}_K\left(\hat{P}_n, P^0\right) \\
&\leq \mathbb{D}_K\left(P_\theta, \hat{P}_n\right) + \mathbb{D}_K\left(\hat{P}_n, P^0\right) \\
&\leq \mathbb{D}_K\left(P_\theta, P^0\right) + 2\mathbb{D}_K\left(\hat{P}_n, P^0\right)
\end{aligned}
$$

A bound in expectation

$$
\begin{aligned}
\forall\theta,\ \mathbb{D}_K\left(P_{\hat{\theta}_n^{MMD}}, P^0\right) &\leq \mathbb{D}_K\left(P_{\hat{\theta}_n^{MMD}}, \hat{P}_n\right) + \mathbb{D}_K\left(\hat{P}_n, P^0\right) \\
&\leq \mathbb{D}_K\left(P_\theta, \hat{P}_n\right) + \mathbb{D}_K\left(\hat{P}_n, P^0\right) \\
&\leq \mathbb{D}_K\left(P_\theta, P^0\right) + 2\mathbb{D}_K\left(\hat{P}_n, P^0\right)
\end{aligned}
$$

Theorem

For any P_0, when $X_1, \dots, X_n$ are i.i.d from P_0,

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

A bound in probability

We can replace the control on the expectation of $\mathbb{D}_K\left(\hat{P}_n, P^0\right)$ by a bound that holds with large probability, thanks to McDiarmid's inequality.

A bound in probability

We can replace the control on the expectation of $\mathbb{D}_K\left(\hat{P}_n, P^0\right)$ by a bound that holds with large probability, thanks to McDiarmid's inequality.

Theorem

For any P_0, when $X_1, \ldots, X_n$ are i.i.d from P_0, with probability at least $1-\delta$,

$$\mathbb{D}_K\left(P_{\hat{\theta}_n}, P^0\right) \leq \inf_{\theta \in \Theta} \mathbb{D}_K\left(P_\theta, P^0\right) + \frac{2 + 2\sqrt{2 \log\left(\frac{1}{\delta}\right)}}{\sqrt{n}}.$$

How to compute $\hat{\theta}_n^{MMD}$?

We actually have

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

How to compute $\hat{\theta}_n^{MMD}$?

We actually have

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

and so

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

that can be approximated by sampling from P_θ.

Short bibliography

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.

Short bibliography

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.

Short bibliography

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*.

Short bibliography

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*.

provided the bound in proba. + asymptotic distribution...

Chérief-Abdellatif, B.-E. and Alquier, P. (2019). Finite Sample Properties of Parametric MMD Estimation : Robustness to Misspecification and Dependence. *Preprint arxiv :1912.05737*.

Chérief-Abdellatif, B.-E. and Alquier, P. (2019). Finite Sample Properties of Parametric MMD Estimation : Robustness to Misspecification and Dependence. *Preprint arxiv :1912.05737*.

the results I will present in the following sections.

Chérief-Abdellatif, B.-E. and Alquier, P. (2019). Finite Sample Properties of Parametric MMD Estimation : Robustness to Misspecification and Dependence. *Preprint arxiv :1912.05737*.

the results I will present in the following sections.

Chérie-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes : Robust Bayesian Estimation via Maximum Mean Discrepancy. *Proceedings of AABI*.

Chérief-Abdellatif, B.-E. and Alquier, P. (2019). Finite Sample Properties of Parametric MMD Estimation : Robustness to Misspecification and Dependence. *Preprint arxiv :1912.05737*.

the results I will present in the following sections.

Chérie-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes : Robust Bayesian Estimation via Maximum Mean Discrepancy. *Proceedings of AABI*.

studies a Bayesian version of this estimator.

Chérief-Abdellatif, B.-E. and Alquier, P. (2019). Finite Sample Properties of Parametric MMD Estimation : Robustness to Misspecification and Dependence. *Preprint arxiv :1912.05737.*

the results I will present in the following sections.

Chérie-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes : Robust Bayesian Estimation via Maximum Mean Discrepancy. *Proceedings of AABI.*

studies a Bayesian version of this estimator.

Alquier, P. and Gerber, M. (2020). Universal Robust Regression via Maximum Mean Discrepancy. *Preprint arxiv :2006.00840.*

Chérief-Abdellatif, B.-E. and Alquier, P. (2019). Finite Sample Properties of Parametric MMD Estimation : Robustness to Misspecification and Dependence. *Preprint arxiv :1912.05737*.

the results I will present in the following sections.

Chérie-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes : Robust Bayesian Estimation via Maximum Mean Discrepancy. *Proceedings of AABI*.

studies a Bayesian version of this estimator.

Alquier, P. and Gerber, M. (2020). Universal Robust Regression via Maximum Mean Discrepancy. *Preprint arxiv :2006.00840*.

application to (linear, logistic, Poisson) regression.

Chérief-Abdellatif, B.-E. and Alquier, P. (2019). Finite Sample Properties of Parametric MMD Estimation : Robustness to Misspecification and Dependence. *Preprint arxiv :1912.05737*.

the results I will present in the following sections.

Chérie-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes : Robust Bayesian Estimation via Maximum Mean Discrepancy. *Proceedings of AABI*.

studies a Bayesian version of this estimator.

Alquier, P. and Gerber, M. (2020). Universal Robust Regression via Maximum Mean Discrepancy. *Preprint arxiv :2006.00840*.

application to (linear, logistic, Poisson) regression.

Alquier, P., Chérief-Abdellatif, B.-E., Derumigny, A. and Fermanian, J.-D. (2020). Estimation of copulas via Maximum Mean Discrepancy. *Preprint arxiv :2009.03017*.

Chérief-Abdellatif, B.-E. and Alquier, P. (2019). Finite Sample Properties of Parametric MMD Estimation : Robustness to Misspecification and Dependence. *Preprint arxiv :1912.05737.*

the results I will present in the following sections.

Chérie-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes : Robust Bayesian Estimation via Maximum Mean Discrepancy. *Proceedings of AABI.*

studies a Bayesian version of this estimator.

Alquier, P. and Gerber, M. (2020). Universal Robust Regression via Maximum Mean Discrepancy. *Preprint arxiv :2006.00840.*

application to (linear, logistic, Poisson) regression.

Alquier, P., Chérief-Abdellatif, B.-E., Derumigny, A. and Fermanian, J.-D. (2020). Estimation of copulas via Maximum Mean Discrepancy. *Preprint arxiv :2009.03017.*

application to copulas + R package : *MMDCopula*.

1. Estimation via MMD optimization
   - Definition of the estimator
   - Basic properties
   - References and further works
2. Robustness to outliers
   - Application to Huber contamination model
   - Example : estimation of the mean of a Gaussian
   - Numerical experiments
3. Robustness to dependence
   - Extension to non-independent observations
   - A (new ?) dependence coefficient
   - Example : auto-regressive observations

Huber contamination model

Reminder

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

Huber contamination model

Reminder

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

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

Huber contamination model

Reminder

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

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

$$\begin{aligned}\mathbb{D}_K(P_{\theta_0}, P_0) &= \|P_{\theta_0} - [(1-\varepsilon)P_{\theta_0} + \varepsilon Q]\|_{\mathcal{H}} \\ &\leq \varepsilon\|P_{\theta_0}\|_{\mathcal{H}} + \varepsilon\|Q\|_{\mathcal{H}} \\ &= 2\varepsilon.\end{aligned}$$

Huber contamination model

Reminder

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

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

$$\mathbb{D}_K(P_{\theta_0}, P_0) \leq 2\varepsilon.$$

Corollary

When $X_1, \dots, X_n$ are i.i.d from $(1-\varepsilon)P_{\theta_0} + \varepsilon Q$,

$$\mathbb{E}\left[\mathbb{D}_K\left(P_{\hat{\theta}_n^{MMD}}, P_{\theta_0}\right)\right] \leq 4\varepsilon + \frac{2}{\sqrt{n}}.$$

Example : Gaussian mean estimation

Example : the model is given by $p_\theta = \mathcal{N}(\theta, \sigma^2 I)$ for $\theta \in \mathbb{R}^d$.

Example : Gaussian mean estimation

Example : the model is given by $p_\theta = \mathcal{N}(\theta, \sigma^2 I)$ for $\theta \in \mathbb{R}^d$.

Using a Gaussian kernel $K(x,y) = \exp(-\|x-y^2\|/\gamma^2)$, from the previous theorem and from the equality

$$\mathbb{D}_K^2 (P_\theta, P_{\theta'}) = 2 \left(\frac{\gamma^2}{4\sigma^2 + \gamma^2} \right)^{\frac{d}{2}} \left[1 - \exp\left(- \frac{\|\theta - \theta'\|^2}{4\sigma^2 + \gamma^2} \right) \right]$$

we obtain

$$\mathbb{E}\left[\|\hat{\theta}_n^{MMD} - \theta_0\|^2 \right] \leq -(4\sigma^2 + \gamma^2) \log \left[1 - 4 \left(\frac{1}{n} + \varepsilon^2 \right) \left(\frac{4\sigma^2 + \gamma^2}{\gamma^2} \right)^{\frac{d}{2}} \right].$$

Example : Gaussian mean estimation

Example : the model is given by $p_\theta = \mathcal{N}(\theta, \sigma^2 I)$ for $\theta \in \mathbb{R}^d$.

Using a Gaussian kernel $K(x,y) = \exp(-\|x - y^2\|/\gamma^2)$, from the previous theorem and from the equality

$$\mathbb{D}_K^2 (P_\theta, P_{\theta'}) = 2 \left(\frac{\gamma^2}{4\sigma^2 + \gamma^2} \right)^{\frac{d}{2}} \left[1 - \exp\left(- \frac{\|\theta - \theta'\|^2}{4\sigma^2 + \gamma^2} \right) \right]$$

we obtain

$$\mathbb{E}\left[\|\hat{\theta}_n^{MMD} - \theta_0\|^2 \right] \quad \text{take } \gamma = 2d\sigma^2$$

$$\leq -(4\sigma^2 + \gamma^2) \log \left[1 - 4 \left(\frac{1}{n} + \varepsilon^2 \right) \left(\frac{4\sigma^2 + \gamma^2}{\gamma^2} \right)^{\frac{d}{2}} \right].$$

Example : Gaussian mean estimation

Example : the model is given by $p_\theta = \mathcal{N}(\theta, \sigma^2 I)$ for $\theta \in \mathbb{R}^d$.

Using a Gaussian kernel $K(x,y) = \exp(-\|x - y^2\|/\gamma^2)$, from the previous theorem and from the equality

$$\mathbb{D}_K^2\left(P_\theta, P_{\theta'}\right) = 2\left(\frac{\gamma^2}{4\sigma^2+\gamma^2}\right)^{\frac{d}{2}}\left[1 - \exp\left(-\frac{\|\theta-\theta'\|^2}{4\sigma^2+\gamma^2}\right)\right]$$

we obtain

$$\mathbb{E}\left[\|\hat{\theta}_n^{MMD} - \theta_0\|^2\right] \lesssim d\sigma^2\left(\frac{1}{n} + \varepsilon^2\right).$$

Example : Gaussian mean estimation, simulations

Model : $\mathcal{N}(\theta, 1)$, and $X_1, \dots, X_n$ i.i.d $\mathcal{N}(\theta_0, 1)$, $n = 100$ and we repeat the experiment 200 times.

	$\hat{\theta}_n^{MLE}$	$\hat{\theta}_n^{MMD}$
mean absolute error	0.0722	0.0838

Example : Gaussian mean estimation, simulations

Model : $\mathcal{N}(\theta, 1)$, and $X_1, \dots, X_n$ i.i.d $\mathcal{N}(\theta_0, 1)$, $n = 100$ and we repeat the experiment 200 times.

	$\hat{\theta}_n^{MLE}$	$\hat{\theta}_n^{MMD}$
mean absolute error	0.0722	0.0838

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

mean absolute error	0.2349	0.0953

Example : Gaussian mean estimation, simulations

Model : $\mathcal{N}(\theta, 1)$, and $X_1, \dots, X_n$ i.i.d $\mathcal{N}(\theta_0, 1)$, $n = 100$ and we repeat the experiment 200 times.

	$\hat{\theta}_n^{MLE}$	$\hat{\theta}_n^{MMD}$
mean absolute error	0.0722	0.0838

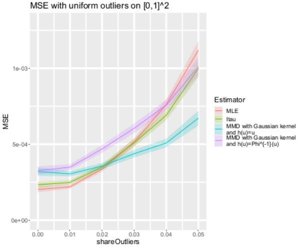
Now, $\varepsilon = 2\%$ of the observations drawn from a Cauchy.

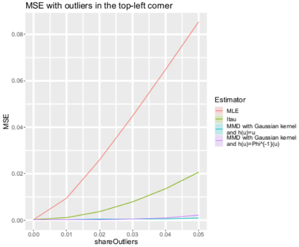
mean absolute error	0.2349	0.0953

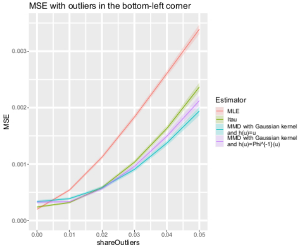
Now, $\varepsilon = 1\%$ are replaced by $1,000$.

mean absolute error	10.018	0.0903

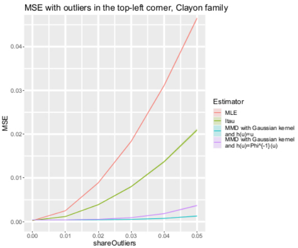
Example : Gaussian copulas

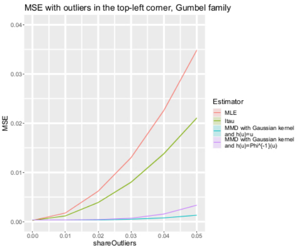


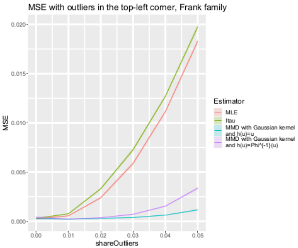




Example :other copula models







1. Estimation via MMD optimization
   - Definition of the estimator
   - Basic properties
   - References and further works

2. Robustness to outliers
   - Application to Huber contamination model
   - Example : estimation of the mean of a Gaussian
   - Numerical experiments

3. Robustness to dependence
   - Extension to non-independent observations
   - A (new ?) dependence coefficient
   - Example : auto-regressive observations

And now, non-independent observations

Lemma

When $X_1, \dots, X_n$ are identically distributed from P_0,

$$\mathbb{E}\left[\mathbb{D}_K\left(\hat{P}_n, P^0\right)\right] \leq ?$$

And now, non-independent observations

Lemma

When $X_1, \dots, X_n$ are identically distributed from P_0,

$$\mathbb{E}\left[\mathbb{D}_K\left(\hat{P}_n, P^0\right)\right] \leq ?$$

$$\begin{aligned}\mathbb{E}\left[\mathbb{D}_K^2\left(\hat{P}_n, P^0\right)\right] &= \mathbb{E}\left[\left\|(1/n)\sum(\mu(\delta_{X_i}) - \mu(P_0))\right\|_{\mathcal{H}}^2\right] \\ &= \frac{1}{n} + \frac{2}{n^2}\sum_{1\leq i<j\leq n}\mathbb{E}\left\langle \mu(\delta_{X_i}) - \mu(P_0), \mu(\delta_{X_j}) - \mu(P_0)\right\rangle_{\mathcal{H}}\end{aligned}$$

Mesure of dependence via covariance in $\mathcal{H}$

Definition

When $(X_1, \dots, X_n, \dots)$ is a stationary process with marginal distribution P_0, we put :

$$\varrho_h = \left| \mathbb{E} \left\langle \mu(\delta_{X_{t+h}}) - \mu(P_0), \mu(\delta_{X_t}) - \mu(P_0) \right\rangle_{\mathcal{H}} \right| .$$

Mesure of dependence via covariance in $\mathcal{H}$

Definition

When $(X_1, \dots, X_n, \dots)$ is a stationary process with marginal distribution P_0, we put :

$$\varrho_h = \left| \mathbb{E} \left\langle \mu(\delta_{X_{t+h}}) - \mu(P_0), \mu(\delta_{X_t}) - \mu(P_0) \right\rangle_{\mathcal{H}} \right| .$$

Lemma - dependent case

When $X_1, \dots, X_n$ are identically distributed from P_0,

$$\mathbb{E}\left[\mathbb{D}_K \left(\hat{P}_n, P^0 \right) \right] \leq \frac{1}{n} \left[1 + \sum_{h=1}^{n} \varrho_h \right]$$

Mesure of dependence via covariance in $\mathcal{H}$

Theorem - dependent case

When $(X_1, \dots, X_n, \dots)$ is a stationary process with marginal distribution P_0

$$\mathbb{E}\left[\mathbb{D}_K\left(P_{\hat{\theta}_n^{MMD}}, P_0\right)\right] \leq \inf_{\theta\in\Theta} \mathbb{D}_K(P_\theta, P_0) + \frac{2 + 2\sum_{h=1}^{n} \varrho_h}{\sqrt{n}}.$$

Mesure of dependence via covariance in $\mathcal{H}$

Theorem - dependent case

When $(X_1, \dots, X_n, \dots)$ is a stationary process with marginal distribution P_0

$$\mathbb{E}\left[\mathbb{D}_K\left(P_{\hat{\theta}_n^{MMD}}, P_0\right)\right] \leq \inf_{\theta\in\Theta} \mathbb{D}_K(P_\theta, P_0) + \frac{2 + 2\sum_{h=1}^{n} \varrho_h}{\sqrt{n}}.$$

1. assume that $\sum_{h=1}^{\infty} \varrho_h = \Sigma < +\infty$ then

$$\mathbb{E}\left[\mathbb{D}_K\left(P_{\hat{\theta}_n^{MMD}}, P_0\right)\right] \leq \inf_{\theta\in\Theta} \mathbb{D}_K(P_\theta, P_0) + \frac{2 + 2\Sigma}{\sqrt{n}}.$$

Mesure of dependence via covariance in $\mathcal{H}$

Theorem - dependent case

When $(X_1, \dots, X_n, \dots)$ is a stationary process with marginal distribution P_0

$$\mathbb{E}\left[\mathbb{D}_K\left(P_{\hat{\theta}_n^{MMD}}, P_0\right)\right] \leq \inf_{\theta\in\Theta} \mathbb{D}_K(P_\theta, P_0) + \frac{2 + 2\sum_{h=1}^{n} \varrho_h}{\sqrt{n}}.$$

1. assume that $\sum_{h=1}^{\infty} \varrho_h = \Sigma < +\infty$ then

$$\mathbb{E}\left[\mathbb{D}_K\left(P_{\hat{\theta}_n^{MMD}}, P_0\right)\right] \leq \inf_{\theta\in\Theta} \mathbb{D}_K(P_\theta, P_0) + \frac{2 + 2\Sigma}{\sqrt{n}}.$$

2. we also have a bound in probability, based on Rio's version of Hoeffding's inequality ; it requires more assumptions.

An example : auto-regressive processes

Proposition

Assume that X_t takes values in $\mathbb{R}^d$ and that $K(x, y) = F(\|x - y\|)$ where F is an L-Lipschitz function. Assume that

$$X_{t+1} = AX_t + \varepsilon_{t+1}$$

where the (ε_t) are i.i.d with $\mathbb{E}\|\varepsilon_0\| < \infty$, and A is a matrix with $\|A\| = \sup_{\|x\|=1} \|Ax\| < 1$.

An example : auto-regressive processes

Proposition

Assume that X_t takes values in $\mathbb{R}^d$ and that $K(x, y) = F(\|x - y\|)$ where F is an L-Lipschitz function. Assume that

$$X_{t+1} = AX_t + \varepsilon_{t+1}$$

where the (ε_t) are i.i.d with $\mathbb{E}\|\varepsilon_0\| < \infty$, and A is a matrix with $\|A\| = \sup_{\|x\|=1} \|Ax\| < 1$.
Then

$$\varrho_t \leq \|A\|^t \frac{2L\mathbb{E}\|\varepsilon_0\|}{1 - \|A\|} \text{ and } \Sigma = \sum_{t=1}^{\infty} \varrho_t = \frac{2\|A\| L\mathbb{E}\|\varepsilon_0\|}{(1 - \|A\|)^2}.$$

A non-mixing process with $\Sigma < +\infty$

Example : consider $X_0 \sim \mathcal{U}([0,1])$, η_t i.i.d $\mathcal{B}e(1/2)$ and

$$X_{t+1} = \frac{X_t + \eta_{t+1}}{2}.$$

A non-mixing process with $\Sigma < +\infty$

Example : consider $X_0 \sim \mathcal{U}([0,1])$, η_t i.i.d $\mathcal{B}e(1/2)$ and

$$X_{t+1} = \frac{X_t + \eta_{t+1}}{2}.$$

It satisfies the assumptions of the previous proposition, we have $\varrho_t \leq L/2^t$ and $\Sigma = 2L$.

A non-mixing process with $\Sigma < +\infty$

Example : consider $X_0 \sim \mathcal{U}([0,1])$, η_t i.i.d $\mathcal{B}e(1/2)$ and

$$X_{t+1} = \frac{X_t + \eta_{t+1}}{2}.$$

It satisfies the assumptions of the previous proposition, we have $\varrho_t \leq L/2^t$ and $\Sigma = 2L$.

Note however that this process is known to be non-mixing.

A non-mixing process with $\Sigma < +\infty$

Example : consider $X_0 \sim \mathcal{U}([0,1])$, η_t i.i.d $\mathcal{B}e(1/2)$ and

$$X_{t+1} = \frac{X_t + \eta_{t+1}}{2}.$$

It satisfies the assumptions of the previous proposition, we have $\varrho_t \leq L/2^t$ and $\Sigma = 2L$.

Note however that this process is known to be non-mixing.

More generally, we prove the following result :

Proposition

Under some (non-restrictive) assumption on the kernel K,

$$\varrho_t \leq c_K.\beta_t \text{ (the } \beta\text{-mixing coef.)}$$

Thank you !