MMD-Bayes

Robust Bayesian Estimation via Maximum Mean Discrepancy

Pierre Alquier

RIKEN

AIP
Center for
Advanced Intelligence Project

AIP PI Seminar – January 10, 2020

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

Please visit the team website

https ://emtiyaz.github.io/

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

Please visit the team website

https ://emtiyaz.github.io/

One of the recurring idea in the team's work :

$$\pi(\theta|X) \propto \underbrace{\pi(X|\theta)}_{\text{likelihood}} \underbrace{\pi(\theta)}_{\textit{prior}} \rightarrow \pi(\theta|X) \propto \underbrace{\exp\left(-\alpha L(X,\theta)\right)}_{\text{loss function}} \underbrace{\pi(\theta)}_{\textit{prior}}.$$

RIKEN AIP Center for Advanced Intelligence Project

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

Please visit the team website

https ://emtiyaz.github.io/

One of the recurring idea in the team's work :

$$\pi(\theta|X) \propto \underbrace{\pi(X|\theta)}_{\text{likelihood}} \underbrace{\pi(\theta)}_{\textit{prior}} \rightarrow \pi(\theta|X) \propto \underbrace{\exp\left(-\alpha L(X,\theta)\right)}_{\text{loss function}} \underbrace{\pi(\theta)}_{\textit{prior}}.$$

P. Alquier, N. Chopin and J. Ridgway (2016). On the Properties of Variational Approximations of Gibbs Posteriors. *Journal of Machine Learning Research*.

K. Osawa, S. Swaroop, A. Jain, R. Eschenhagen, R. E. Turner, R. Yokota, M. E. Khan (2019). Practical Deep Learning with Bayesian Principles. *NeurIPS*.

Today : a loss function leading to robust estimation.
Joint works with :

Badr-Eddine Chérief-Abdellatif.

ENSAE
IP PARIS

B.-E. Chérief-Abdellatif, P. Alquier (2019). MMD-Bayes : Robust Bayesian Estimation via Maximum Mean Discrepancy. AABI 2019.

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

1. Introduction : some problems with the likelihood
2. Kernels and MMD distance
   - Kernels and RKHS
   - Maximum Mean Discrepancy (MMD)
   - MMD-based estimation
3. MMD-Bayes
   - Bayesian estimator based on MMD
   - Consistency and rate of convergence
   - Simulation study

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

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

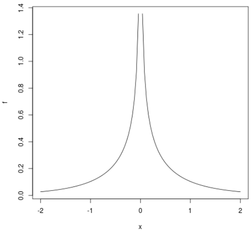
Example : $P_{(m,\sigma)} = \mathcal{N}(m, \sigma^2)$ then

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

MLE not unique / not consistent

Example :

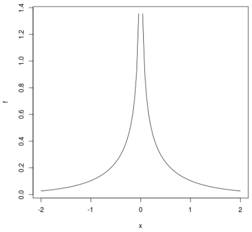
$$p_\theta(x) = \frac{\exp(-|x-\theta|)}{2\sqrt{\pi|x-\theta|}},$$



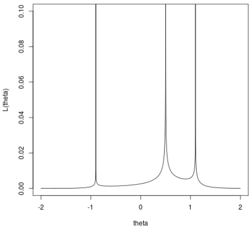
MLE not unique / not consistent

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

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

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

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

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

$$P_0 = (1-\varepsilon).\mathcal{U}nif[0,1] + \varepsilon.\mathcal{N}(10034,1)...$$

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 ! ! !

1 Introduction : some problems with the likelihood

2 Kernels and MMD distance
- Kernels and RKHS
- Maximum Mean Discrepancy (MMD)
- MMD-based estimation

3 MMD-Bayes
- Bayesian estimator based on MMD
- Consistency and rate of convergence
- Simulation study

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 $(x_1, \dots, x_n) \in \mathcal{X}^n$ and $(c_1, \dots, c_n) \neq (0, \dots, 0) \in \mathbb{R}^n$,

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

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 $\mathcal{X}$,

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

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

Definition - MMD based estimator

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

MMD-based estimator

Reminder of the context : $X_1, \dots, X_n$ be i.i.d in $\mathcal{X}$ from a probability distribution P_0, model $(P_\theta, \theta \in \Theta)$.

Definition - MMD based estimator

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

Even though this idea was sometimes used before, the first theoretical study :

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

Universal estimation with MMD

Theorem - Chérief-Abdellatif + PA

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

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

Universal estimation with MMD

Theorem - Chérief-Abdellatif + PA

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

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

Corollary - Huber contamination model

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

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

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.

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

Going beyond toy examples

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

Li, Y., Swersky, K. & Zemel, R. (2015). Generative moment matching networks. *ICML 2015*.

From the first reference :

MMD-Bayes

Given a prior $\pi(\theta)$ we propose the following pseudo-posterior :

$$\pi_\alpha(\theta|X_1,\dots,X_n) \propto \mathrm{e}^{-\alpha\mathbb{D}_K^2\left(P_\theta,\hat{P}_n\right)}\pi(\theta).$$

MMD-Bayes

Given a prior $\pi(\theta)$ we propose the following pseudo-posterior :

$$\pi_\alpha(\theta|X_1,\dots,X_n) \propto \mathrm{e}^{-\alpha\mathbb{D}^2_K\left(P_\theta,\hat{P}_n\right)}\pi(\theta).$$

We also define a variational approximation for this. Given a set $\mathcal{F}$ of probability distributions,

$$\hat{\pi}_\alpha(\theta) = \arg\min_{q\in\mathcal{F}} \mathcal{K}[q,\pi_\alpha(\cdot|X_1,\dots,X_n)].$$

Bayesian MMD-based estimation

Theorem - Chérief-Abdellatif + PA

Let $\mathcal{B} = \{\theta \in \Theta / \mathbb{D}_K\left(P_{\theta_0}, P_\theta\right) \leq 1/\sqrt{n}\}$. Assume (π, α) satisfies the prior mass condition : $\pi(\mathcal{B}) \geq e^{-\alpha/\sqrt{n}}$. Then :

$$\mathbb{E}\left[\int \mathbb{D}_K\left(P_\theta, P^0\right) \pi_n^\beta(\mathrm{d}\theta)\right] \leq 4 \inf_{\theta \in \Theta} \mathbb{D}_K\left(P_\theta, P^0\right) + \frac{4}{\sqrt{n}}.$$

Bayesian MMD-based estimation

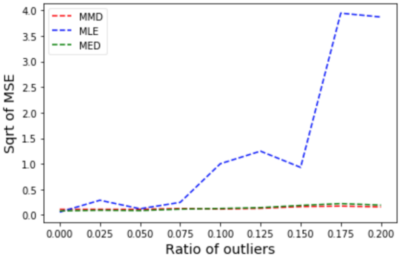
Theorem - Chérief-Abdellatif + PA

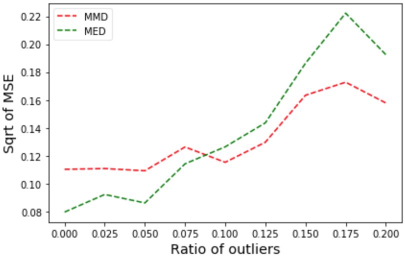
Let $\mathcal{B} = \{\theta \in \Theta / \mathbb{D}_K(P_{\theta_0}, P_\theta) \leq 1/\sqrt{n}\}$. Assume (π, α) satisfies the prior mass condition : $\pi(\mathcal{B}) \geq e^{-\alpha/\sqrt{n}}$. Then :

$$\mathbb{E}\left[\int \mathbb{D}_K\left(P_\theta, P^0\right) \pi_n^\beta(\mathrm{d}\theta)\right] \leq 4 \inf_{\theta\in\Theta} \mathbb{D}_K\left(P_\theta, P^0\right) + \frac{4}{\sqrt{n}}.$$

A similar result holds for the variational approximation.

Experiments in the Gaussian model





Thank you !