The nice properties of MMD for statistical estimation

Pierre Alquier

RIKEN

AIP
Center for
Advanced Intelligence Project

MAD-stat seminar, Toulouse School of Economics. – Nov. 10, 2022

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

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

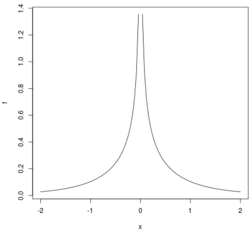
Example : $P_{(m,\sigma)} = \mathcal{N}(m, \sigma^2)$ then

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

MLE not unique / not consistent

Example :

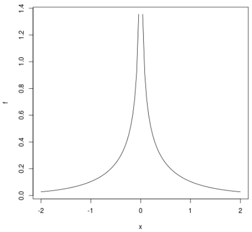
$$p_\theta(x) = \frac{\exp(-|x-\theta|)}{2\sqrt{\pi|x-\theta|}},$$



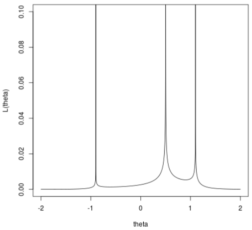
MLE not unique / not consistent

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 :

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

Contents

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

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

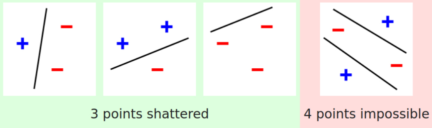
where $\epsilon_1, \dots, \epsilon_n$ are i.i.d Rademacher variables :

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

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 $\mathcal{A} \subseteq \mathcal{P}(\mathcal{X})$,

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

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,
  - $\mathrm{VC}(\mathcal{F}) = 1$, so :

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

Example 2 : Maximum Mean Discrepancy (MMD)

- RKHS $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ with kernel $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.
- If $\|\phi(x)\|_{\mathcal{H}} = k(x, x) \leq 1$ then $\mathbb{E}_{X \sim \mu}[\phi(X)]$ is well-defined .
- The map $\mu \mapsto \mathbb{E}_{X \sim \mu}[\phi(X)]$ is one-to-one if k is characteristic.
- Gaussian kernel $k(x, y) = \exp(-\|x - y\|^2/\gamma^2)$ satisfies these assumption.

$$\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}.$$

$$\begin{aligned}\mathbb{D}_k(P, Q) := d_{\mathcal{F}}(P, Q) &= \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X \sim P}[f(X)] - \mathbb{E}_{X \sim Q}[f(X)] \right| \\ &= \left\| \mathbb{E}_{X \sim P}[\phi(X)] - \mathbb{E}_{X \sim Q}[\phi(X)] \right\|_{\mathcal{H}}.\end{aligned}$$

Example 2 : MMD

Theorem (Bartlett and Mendelson)

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

Corollary

$$\mathbb{E}\left[\mathbb{D}_k(P_{\hat{\theta}_{\mathbb{D}_k}}, P_0)\right] \leq \inf_{\theta \in \Theta} \mathbb{D}_k(P_\theta, P_0) + 4\sqrt{\frac{\sup_x k(x,x)}{n}}.$$

Example 2 : 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 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 $\mathrm{Lip}(f) := \sup_{x\neq y} |f(x) - f(y)|/\delta(x,y)$.

Bound on the Rademacher complexity when $\mathcal{X}$ is bounded :

Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Schölkopf, B., Lanckriet, G.R. (2010). Non-parametric estimation of integral probability metrics. IEEE International Symposium on Information Theory.

Minimum Wasserstein estimation studied in :

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

Improving the constant

From now, we assume that $\sup_x k(x,x) \leq 1$. We know :

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

We will now prove a better result without using the Rademacher complexity :

Theorem

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

Proof of the theorem : preliminary lemma

Lemma

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

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

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

Proof of the theorem

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

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

A bound in probability

Thanks to McDiarmid's inequality :

Theorem

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

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

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

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

Example : Gaussian mean estimation

Example : $P_\theta = \mathcal{N}(\theta, \sigma^2 I)$ for $\theta \in \mathbb{R}^d$.
Using a Gaussian kernel $k(x, y) = \exp(-\|x - y^2\|/\gamma^2)$,

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

Together with the previous result, this gives :

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

$\gamma = 2d\sigma^2 \Rightarrow$

$$\|\hat{\theta}_n^{MMD} - \theta_0\|^2 \leq d\sigma^2 \frac{8e(1 + \sqrt{2\log 1/\delta})^2}{n}(1 + o(1)).$$

Variance-aware bounds (1/2)

$$\left\{\mathbb{E}\left[\mathbb{D}_k\left(\hat{P}_n, P^0\right)\right]\right\}^2 \leq \mathbb{E}\left[\mathbb{D}_k^2\left(\hat{P}_n, P^0\right)\right]$$
$$= \mathbb{E}\left[\left\|(1/n)\sum(\mu(\delta_{X_i}) - \mu(P_0))\right\|_{\mathcal{H}}^2\right]$$
$$= (1/n)\underbrace{\mathbb{E}\left[\|\mu(\delta_{X_1}) - \mu(P_0)\|_{\mathcal{H}}^2\right]}_{=: v_k(P_0)}$$

Lemma - variance-aware version

$$\mathbb{E}\left[\mathbb{D}_k\left(\hat{P}_n, P^0\right)\right] \leq \sqrt{\frac{v_k(P_0)}{n}} \leq \sqrt{\frac{1}{n}}.$$

Variance-aware bounds (2/2)

Theorem – bound in expectation

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

Theorem – bound in probability

With probability at least $1-\delta$,

$$\mathbb{D}_k\left(P_{\hat{\theta}}, P^0\right) \leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, P^0\right) + 2\sqrt{\frac{v_k(P_0) 2\log\frac{1}{\delta}}{n}} + \frac{8\log\frac{1}{\delta}}{3n}.$$

Joint work with Geoffrey Wolfer (RIKEN AIP).

Wolfer, G. and Alquier, P. Variance-Aware Estimation of Kernel Mean Embedding. Preprint arXiv :2210.06672.

Upper-bounding the variance $v_k(P_0)$

In the case of the Gaussian kernel

$$k(x,y) = \exp(-\|x-y\|^2/\gamma^2)$$

we have

$$v_k(P_0) \leq 1 - \exp\left[-\frac{2\mathrm{Tr}(Var_{P_0}(X))}{\gamma^2}\right] \leq \begin{cases} \frac{2\mathrm{Tr}(Var_{P_0}(X))}{\gamma^2} \\ 1. \end{cases}$$

Example : Gaussian mean estimation (continued).

Using the variance aware bound

$$\gamma = \gamma_n \to +\infty \Rightarrow \|\hat{\theta}_n^{MMD} - \theta_0\|^2 \leq d\sigma^2 \frac{4\log 1/\delta}{n}(1+o(1)).$$

Empirical bound

In practice, we can estimate $v_k(P_0)$ by

$$\hat{v}_k := \frac{1}{n-1}\sum_{i=1}^{n}\left(k(X_i,X_i) - \frac{1}{n}\sum_{j=1}^{n} k(X_i,X_j)\right).$$

We have $\mathbb{E}(\hat{v}_k) = v_k(P_0)$, and

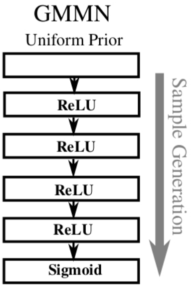
Theorem – bound with empirical variance

Assume that $k(x,y) = \psi(x-y) \in [a,b]$. Then, with probability at least $1-\delta$,

$$\mathbb{D}_k\left(P_{\hat{\theta}}, P^0\right) \leq \inf_{\theta\in\Theta} \mathbb{D}_k\left(P_\theta, P^0\right) + 2\sqrt{\frac{\hat{v}_k 2\log\frac{1}{\delta}}{n}} + \frac{32\sqrt{b-a}\log\frac{1}{\delta}}{3n}.$$

Contents

Generative Adversarial Networks (GAN, 1/2)



Generative model $X \sim P_\theta$:

- $U \sim \text{Unif}[0,1]^d$,
- $X = F_\theta(U)$ where F_θ is some NN with weights θ.

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

Li, Y., Swersky, K. & Zemel, R. (2015). Generative Moment Matching Networks. ICML.

$\rightarrow$ proposed to minimize the MMD to learn θ.

GAN (2/2)

Results from Dziugaite et al. (2015).

Inference for Systems of SDEs (1/2)

This paper developped the asymptotic theory of MMD :

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

They also applied the method to inference in SDEs :

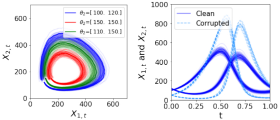
$$\mathrm{d}X_t = b(X_t, \theta_1)\mathrm{d}t + \sigma(X_t, \theta_2)\mathrm{d}W_t$$

- easy to sample from the model with a given $\theta = (\theta_1, \theta_2)$,
- they propose a method to approximate the gradient of the MMD criterion.

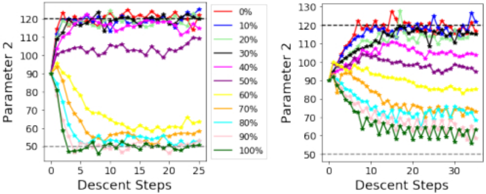
Inference for Systems of SDEs (2/2)

Example in a (stochastic) Lotka-Volterra model.

$$\mathrm{d}\begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} = \left[\begin{pmatrix} 1 \\ 0 \end{pmatrix}\theta_{11}X_{1,t} + \begin{pmatrix} -1 \\ 1 \end{pmatrix}\theta_{12}X_{1,t}X_{2,t} + \begin{pmatrix} 0 \\ -1 \end{pmatrix}\theta_{13}X_{2,t}\right]\mathrm{d}t + \begin{pmatrix} 1 \\ 0 \end{pmatrix}\sqrt{\theta_{11}X_{1,t}}\mathrm{d}W_t^{(1)} + \begin{pmatrix} -1 \\ 1 \end{pmatrix}\sqrt{\theta_{12}X_{1,t}X_{2,t}}\mathrm{d}W_t^{(2)} + \begin{pmatrix} 0 \\ -1 \end{pmatrix}\sqrt{\theta_{13}X_{2,t}}\mathrm{d}W_t^{(3)},$$



Results from Briol et al. (2019) : compare MMD minimization to Wasserstein minimization.



Regression

- problem with regression : we want to specify and estimate a parametric model $P_{\theta(X)}$ for $Y|X$. MMD requires to specify a model for (X, Y).
- natural idea : estimate the distribution of X by $\frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ and use the MMD procedure on $P_{\theta(X)}$.
- the previous theory shows directly that we estimate the distribution of (X, Y) consistently.
- it is far more difficult to prove that we estimate the distribution of $Y|X$.

Joint work with M. Gerber (Bristol).

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

Copulas

- another semi-parametric model : copulas.
- asymptotic theory + R package.


With B.-E. Chérief-Abdellatif (CNRS), J.-D. Fermanian (ENSAE Paris), A. Derumigny (TU Delft).

Alquier, P., Chérief-Abdellatif, B.-E., Derumigny, A. and Fermanian, J.-D. Estimation of copulas via Maximum Mean Discrepancy. JASA, to appear.

CRAN
Mirrors
What's new?
Task Views
Search

About R
R Homepage
The R Journal

Software
R Sources
R Binaries
Packages
Other

Documentation
Manuals
FAQs
Contributed

MMDCopula: Robust Estimation of Copulas by Maximum Mean Discrepancy

Provides functions for the robust estimation of parametric families of copulas using minimization of the Maximum Mean Discrepancy, following the article Alquier, Chérief-Abdellatif, Derumigny and Fermanian (2020) <arXiv:2010.00408>.

Version:	0.1.0
Depends:	R (≥ 3.6.0)
Imports:	VineCopula, cubature, pcaPP, randtoolbox
Suggests:	knitr, rmarkdown
Published:	2020-10-13
Author:	Alexis Derumigny [aut, cre], Pierre Alquier [aut], Jean-David Fermanian [aut], Badr-Eddine Chérief-Abdellatif [aut]
Maintainer:	Alexis Derumigny <a.f.f.derumigny at utwente.nl>
BugReports:	https://github.com/AlexisDerumigny/MMDCopula/issues
License:	GPL-3
NeedsCompilation:	no
Materials:	README NEWS
CRAN checks:	MMDCopula results

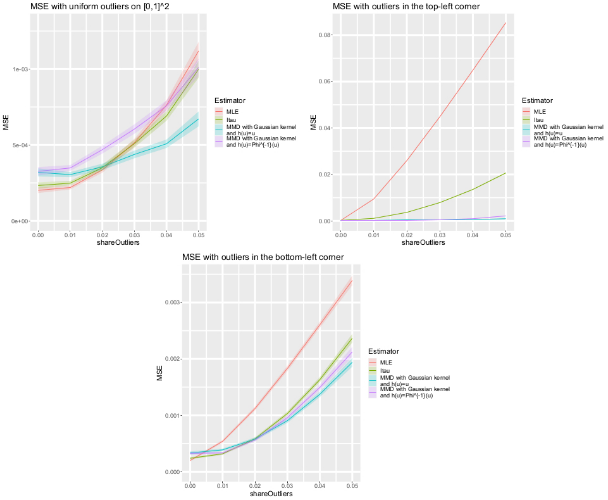
Downloads:

Reference manual:	MMDCopula.pdf
Vignettes:	The MMD copula package: robust estimation of parametric copula models by MMD minimization
Package source:	MMDCopula_0.1.0.tar.gz
Windows binaries:	r-devel: MMDCopula_0.1.0.zip, r-release: MMDCopula_0.1.0.zip, r-oldrel: MMDCopula_0.1.0.zip
macOS binaries:	r-release: MMDCopula_0.1.0.tgz, r-oldrel: MMDCopula_0.1.0.tgz

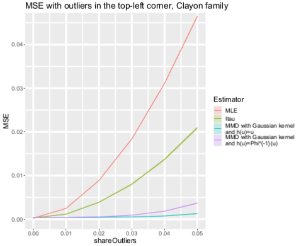
Linking:

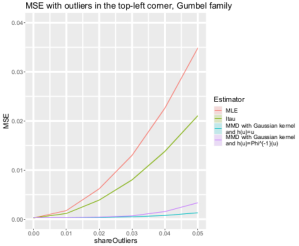
Please use the canonical form https://CRAN.R-project.org/package=MMDCopula to link to this page.

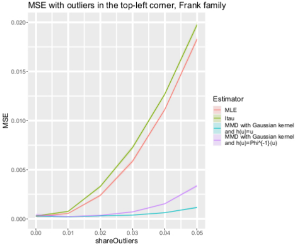
Example : Gaussian copulas



Example : other models





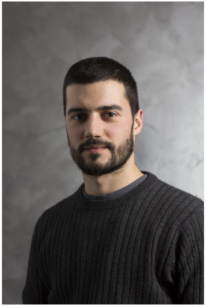


Contents

Co-authors and paper

S. Legramanti, D. Durante & P. Alquier (2022). Concentration and robustness of discrepancy–based ABC via Rademacher complexity. Preprint arXiv :2206.06991.

Sirio Legramanti
(University of Bergamo)



Daniele Durante
(Bocconi University, Milan)

Estimators, randomized estimators and Bayes rule

- $Y_{1:n} = Y_1, \ldots, Y_n$ i.i.d from μ^*,
- model : $(\mu_\theta, \theta \in \Theta)$,
- estimator : $\hat{\theta} = \hat{\theta}(Y_{1:n})$,
- randomized estimator : $\hat{\rho}(\cdot) = \hat{\rho}(Y_{1:n})(\cdot)$ probability measure on Θ.

Examples of randomized estimators :

- posterior : $\hat{\rho}(\theta) = \pi(\theta | Y_{1:n}) \propto \underbrace{\mathcal{L}(\theta; Y_{1:n})}_{\text{likelihood}} \underbrace{\pi(\theta)}_{\text{prior}}$,
- fractional/tempered posterior : $\hat{\rho}(\theta) \propto [\mathcal{L}(\theta; Y_{1:n})]^\alpha \pi(\theta)$,
- Gibbs estimator : $\hat{\rho}(\theta) \propto \exp[-\eta \underbrace{R(\theta; Y_{1:n})}_{\text{loss}}] \pi(\theta)$.

Evaluating randomized estimators

Assume in this slide that $\mu^* = \mu_{\theta_0}$: "the truth is in the model".
Statistical performance of an estimator :

- consistency : $d(\hat{\theta}, \theta_0) \xrightarrow[n\to\infty]{} 0$ (in proba., a.s., ...) ?
- rate of convergence : $\mathbb{E}_{Y_{1:n}}[d(\hat{\theta}, \theta_0)] \leq r_n \xrightarrow[n\to\infty]{} 0$?
- ...

For a randomized estimator :

- contraction rate :

$$\mathbb{P}_{\theta\sim\hat{\rho}}[d(\theta, \theta_0) \geq r_n] \xrightarrow[n\to\infty]{} 0 \text{ (in proba., a.s., ...) ?}$$

- average risk : $\mathbb{E}_{Y_{1:n}}\left[\mathbb{E}_{\theta\sim\hat{\rho}}[d(\theta, \theta_0)]\right] \leq r_n$?
- ...

Approximate Bayesian Inference

- Well-known conditions to prove contraction of the posterior,
- tools from ML for randomized estimators : PAC-Bayes bounds.

Given a "non-exact" algorithm targetting $\hat{\rho}$ instead of $\pi(\cdot|Y_{1:n})$: variational approximations, ABC, etc., we can

- quantify how well $\hat{\rho}$ approximates $\pi(\cdot|Y_{1:n})$?
- study $\hat{\rho}$ as a randomized estimator and study its contraction/convergence.

Reminder on ABC

Approximate Bayesian Computation (ABC)

INPUT : sample $Y_{1:n} = (Y_1, \ldots, Y_n)$, model $(\mu_\theta, \theta \in \Theta)$, prior π, statistic S, metric δ and threshold ϵ.

(i) sample $\theta \sim \pi$,

(ii) sample $Z_{1:n} = (Z_1, \ldots, Z_n)$ i.i.d. from μ_θ :

- if $\delta(S(Y_{1:n}), S(Z_{1:n})) \leq \epsilon$ return θ,
- else goto (i).

OUTPUT : $\vartheta \sim \hat{\rho}$.

- discrete sample space, if S =identity and $\epsilon = 0$, ABC is actually exact : $\hat{\rho}(\cdot) = \pi(\cdot|Y_{1:n})$.
- general case : ABC not exact, we can ask two questions :
  1. is $\hat{\rho}(\cdot)$ a good approximation of $\pi(\cdot|Y_{1:n})$?
  2. is $\hat{\rho}$ a good randomized estimator ?

Discrepancy-based ABC

Approximate Bayesian Computation (ABC)

INPUT : sample $Y_{1:n}$, model $(\mu_\theta, \theta \in \Theta)$, prior π, IPM $d_{\mathcal{F}}$ and threshold ϵ.

(i) sample $\theta \sim \pi$,

(ii) sample $Z_{1:n}$ i.i.d. from μ_θ :

- if $d_{\mathcal{F}}(\hat{\mu}_{Y_{1:n}}, \hat{\mu}_{Z_{1:n}}) \leq \epsilon$ return θ,
- else goto (i).

OUTPUT : $\vartheta \sim \hat{\rho}_\epsilon$.

Rermark : when $d_{\mathcal{F}}$ is the MMD with kernel k,

$$d_{\mathcal{F}}(\hat{\mu}_{Y_{1:n}}, \hat{\mu}_{Z_{1:n}}) = \sum_{i,j} k(Y_i, Y_j) - 2 \sum_{i,j} k(Y_i, Z_j) + \sum_{i,j} k(Z_i, Z_j).$$

Approximation of the posterior

Bernton, E., Jacob, P. E., Gerber, M. & Robert, C. P. (2019). Approximate Bayesian Computation with the Wasserstein distance. JRSS-B.

Contains a general result that can be applied here.

Theorem

Assume

- μ_θ has a continuous density f_θ and for some neighborhood V of $Y_{1:n}$ we have $\sup_{\theta\in\Theta} \sup_{v_{1:n}\in V} \prod_{i=1}^{n} f_\theta(v_i) < +\infty$.
- $v_{1:n} \mapsto d_{\mathcal{F}}(\hat{\mu}_{Y_{1:n}}, \hat{\mu}_{v_{1:n}})$ is continuous.

Then

$$\forall \text{ measurable set } A,\ \hat{\rho}_\epsilon(A) \xrightarrow[\epsilon\to 0]{} \pi(A|Y_{1:n}).$$

Assumptions for contraction

(C1) $\mathcal{Y}$-valued $Y_{1:n} = (Y_1, \dots, Y_n)$ i.i.d from μ_*, put :

$$\epsilon^* := \inf_{\theta\in\Theta} d_{\mathcal{F}}(\mu_\theta, \mu_*).$$

(C2) prior mass condition : there is $c > 0, L \geq 1$ such that

$$\pi\left(\left\{\theta \in \Theta : d_{\mathcal{F}}(\mu_\theta, \mu_*) - \epsilon^* \leq \epsilon\right\}\right) \geq c\epsilon^L$$

(C3) functions in $\mathcal{F}$ are bounded :

$$\sup_{f\in\mathcal{F}} \sup_{y\in\mathcal{Y}} |f(y)| \leq b.$$

(C4) the Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$ satisfies

$$\mathfrak{R}_n(\mathcal{F}) \xrightarrow[n\to\infty]{} 0.$$

Contraction of discrepancy-based ABC

Theorem 1

Under (C1)-(C4), with $\epsilon := \epsilon_n = \epsilon^* + \bar{\epsilon}_n$ with $\bar{\epsilon}_n \to 0$, $n\bar{\epsilon}_n^2 \to \infty$ and $\bar{\epsilon}_n / \mathfrak{R}_n(\mathcal{F}) \to \infty$. Then, for any sequence $M_n > 1$,

$$\hat{\rho}_{\epsilon_n}\left(\left\{\theta \in \Theta : d_{\mathcal{F}}(\mu_\theta, \mu_*) > \epsilon^* + r_n\right\}\right) \leq \frac{2 \cdot 3^L}{cM_n}$$

$$\text{where } r_n = \frac{4\bar{\epsilon}_n}{3} + 2\mathfrak{R}_n(\mathfrak{F}) + b\sqrt{\frac{2\log(\frac{M_n}{\bar{\varepsilon}_n^L})}{n}},$$

with probability $\to 1$ with respect to the sample $Y_{1:n}$.

Examples

- Assume $\mathfrak{R}_n(\mathcal{F}) \leq c\sqrt{1/n}$ (MMD, Kolmogorov...). Take $M_n = n$ and $\bar{\epsilon}_n = \sqrt{\log(n)/n}$ to get

$$\hat{\rho}_{\epsilon_n}\left(\left\{\theta \in \Theta : d_{\mathcal{F}}(\mu_\theta, \mu_*) > \epsilon^* + r_n\right\}\right) \leq \frac{2 \cdot 3^L}{cn}$$

where $r_n = \mathcal{O}\left(\sqrt{\log(n)/n}\right)$.

- Larger $\mathfrak{R}_n(\mathcal{F})$ will lead to slower rates.

Removing (C3)-(C4)

- if we remove (C3)-(C4), we cannot use classical concentration results on $d_{\mathcal{F}}\left(\mu_*, \hat{\mu}_{Y_{1:n}}\right)$ and $d_{\mathcal{F}}\left(\mu_\theta, \hat{\mu}_{Z_{1:n}}\right)$.
- we can still provide a result under the assumption that "some concentration holds", as

Bernton, E., Jacob, P. E., Gerber, M. & Robert, C. P. (2019). Approximate Bayesian Computation with the Wasserstein distance. JRSS-B.

for the Wasserstein distance.

- however, this will impose assumptions on μ_*, $\{\mu_\theta, \theta \in \Theta\}$ and might lead to slower contraction rates. In our paper, we illustrate this with MMD with unbounded kernels :

$$\mathfrak{R}_n(\mathcal{F}) \leq \sqrt{\frac{\sup_{y\in\mathcal{Y}} k(y,y)}{n}} = +\infty.$$

Example : MMD-ABC with unbounded kernel

Theorem 2

Under (C1)-(C2), and

(C5) $\mathbb{E}_{Y\sim\mu_*}[k(Y,Y)] < +\infty$,

(C6) $\sup_{\theta\in\Theta}\mathbb{E}_{Z\sim\mu_\theta}[k(Z,Z)] < +\infty$,

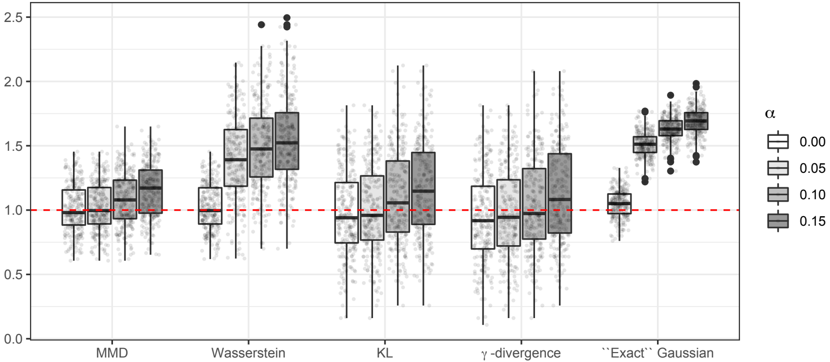
$\epsilon_n = \epsilon^* + \bar{\epsilon}_n$ with $\bar{\epsilon}_n \to 0$. Then, for some $C > 0$, for any sequence $M_n > 1$, with proba. $\to 1$,

$$\hat{\rho}_{\epsilon_n}\left(\left\{\theta\in\Theta : d_{\mathcal{F}}(\mu_\theta,\mu_*) > \epsilon^* + r_n\right\}\right) \leq \frac{C}{M_n}$$

$$\text{where } r_n = \frac{4\bar{\epsilon}_n}{3} + \frac{M_n^2}{n^2\bar{\epsilon}^{2L}}.$$

For example $M_n = \sqrt{n}$ we can get $r_n = \mathcal{O}(1/n^{2L+1})$.

Experiments in the Gaussian case



La fin

終わり

ありがとう ございます。