Robust estimation via Minimum Distance Estimation

Pierre Alquier

RIKEN

AIP
Center for
Advanced Intelligence Project

Statistical Estimation and Deep Learning in UQ for PDEs
Vienna, May 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

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

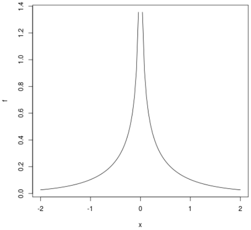
Example : $P_{(m,\sigma)} = \mathcal{N}(m, \sigma^2)$ then

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

MLE not unique / not consistent

Example :

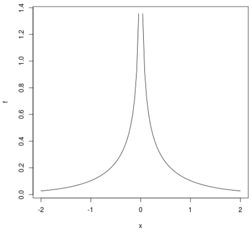
$$p_\theta(x) = \frac{\exp(-|x-\theta|)}{2\sqrt{\pi|x-\theta|}},$$



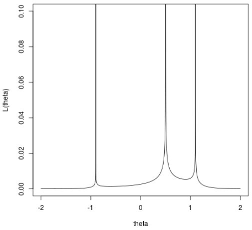
MLE not unique / not consistent

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

Minimum Distance Estimation (MDE)

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

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

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

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

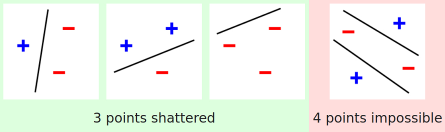


Image from Wikipedia.

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

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

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 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 : Huber's contamination

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

Reminder

$k(x,x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$, k bounded by 1 and characteristic.

Reminder - MMD

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

More explicit formulas for the MMD

We actually have

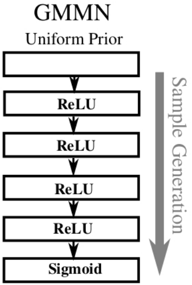
$$\mathrm{MMD}_k^2(P_\theta, \hat{P}_n) = \mathbb{E}_{X,X'\sim P_\theta}[k(X,X')] - \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_{X\sim P_\theta}[k(X_i,X)] + \frac{1}{n^2}\sum_{1\leq i,j\leq n} k(X_i,X_j)$$

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.

Generative Adversarial Networks (GAN, 1/2)



Generative model $X \sim P_\theta$:

- $U \sim \mathrm{Unif}[0,1]^d$,
- $X = F_\theta(U)$ 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)

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

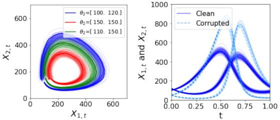
$$\mathrm{d}X_t = b(X_t, \theta_1)\mathrm{d}t + \sigma(X_t, \theta_2)\mathrm{d}W_t$$

- 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.

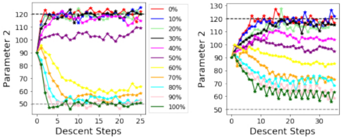
Inference for Systems of SDEs (2/2)

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, MMD approach can be adapted.
-

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

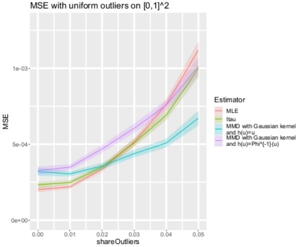
Linking:

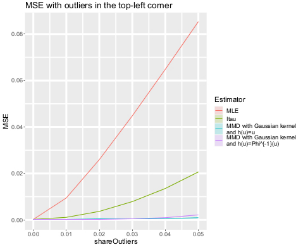
Please use the canonical form https://CRAN.R-project.org/package=MMDCopula to link to this page.

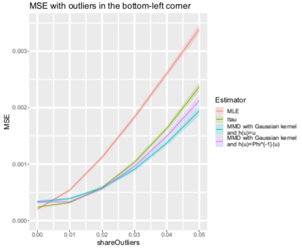
With B.-E. Chérief-Abdellatif (Oxford), J.-D. Fermanian (ENSAE Paris), A. Derumigny (TU Delft).

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

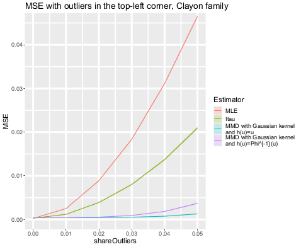
Example : Gaussian copulas

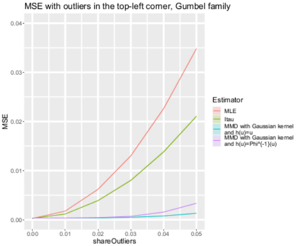


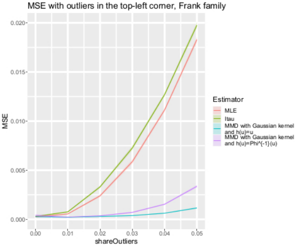




Example : other models







Contents

Generalized posteriors

Posterior

$$\pi(\theta|X_1, \dots, X_n) \propto L_n(\theta)\pi(\theta).$$

Generalized posterior

$$\hat{\pi}_{\beta,R_n}(\theta) \propto \exp(-\beta.R_n(\theta))\pi(\theta).$$

- old idea in ML (PAC-Bayes, forecasting with expert advice...) and in statistics (Gibbs posteriors...)
- popularized / extended and studied by :

Bissiri, P. G., Holmes, C. C. & Walker, S. G. (2016). A general framework for updating belief distributions. *JRSS-B*.

Knoblauch, J., Jewson, J. & Damoulas, T. (2022). An Optimization-centric View on Bayes' Rule : Reviewing and Generalizing Variational Inference. *JMLR* (to appear).

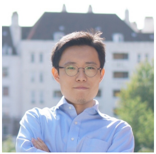
Generalizing the posterior with IPS

Generalized posterior with IPS

$$\hat{\pi}_{\beta,R_n}(\theta) \propto \exp(-\beta . d_{\mathcal{F}}(P_\theta, \hat{P}_n))\pi(\theta).$$

- in the MMD case : non-asymptotic result in

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



- asymptotic results to come very soon in a joint paper with Takuo Matsubara (Newcastle) and Jeremias Knoblauch (UCL).
- both papers discuss computation via variational approximations or MCMC.

Reminder on ABC

Approximate Bayesian Computation (ABC)

input : sample $X_1^n = (X_1, \dots, X_n)$, model $(P_\theta, \theta \in \Theta)$, prior π, statistic S, distance δ and threshold ϵ.

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

(ii) sample $Y_1^n = (Y_1, \dots, Y_n)$ i.i.d. from P_θ :
- if $\delta(S(X_1^n), S(Y_1^n)) \leq \epsilon$ return θ,
- else goto (i).

- how close is the distribution of the output to the posterior $\pi(\theta|X_1, \dots, X_n)$?
- reverse point of view : what are the properties of the "generalized posterior" we sample from ?

ABC with IPS

Here, we study the situation :

- $S(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ the empirical distribution,
- $\delta(P, Q) = d_{\mathcal{F}}(P, Q)$.

IPS–ABC

input : sample $X_1^n = (X_1, \ldots, X_n)$, model $(P_\theta, \theta \in \Theta)$, prior π, set of functions $\mathcal{F}$ and threshold ϵ. Put $\hat{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$.

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

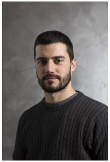
(ii) sample $Y_1^n = (Y_1, \ldots, Y_n)$ i.i.d. from P_θ and put $\hat{P}_n^Y = \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}$,

- if $d_{\mathcal{F}}(\hat{P}_n, \hat{P}_n^Y) \leq \epsilon$ return θ,
- else goto (i).

Notation : the output $\vartheta \sim \hat{\pi}_{n,\epsilon}^{\mathcal{F}}(\cdot)$.

Properties of $\hat{\pi}^{\mathcal{F}}_{n,\epsilon}(\cdot)$

In a forthcoming joint paper with S. Legramanti (University of Bergamo) and D. Durante (Bocconi University, Milan) we study 3 questions :



1. $\hat{\pi}^{\mathcal{F}}_{n,\epsilon}(\theta) \xrightarrow[\epsilon \searrow ?]{?} \pi(\theta|X_1^n)$.
2. $\hat{\pi}^{\mathcal{F}}_{n,\epsilon}(\theta) \xrightarrow[n \to \infty]{?} ?$
3. $\hat{\pi}^{\mathcal{F}}_{n,\epsilon_n}(\cdot) \xrightarrow[n \to \infty]{?} \delta_{\theta_0}$ if $P_0 = P_{\theta_0}$.

Contraction of the ABC posterior

$$\epsilon_* := \inf_{\theta \in \Theta} d_{\mathcal{F}}(P_\theta, P_0).$$

Theorem 3

Assume :

- for all $\epsilon > 0$, $\pi(\{\theta : d_{\mathcal{F}}(P_\theta, P_0) \leq \epsilon_* + \epsilon\}) \geq c\epsilon^d$.
- $\forall f \in \mathcal{F}$, $\sup_{x \in \mathcal{X}} |f(x)| \leq 1$.
- $\mathrm{Rad}_n(\mathcal{F}) \xrightarrow[n \to \infty]{} 0$.

Let ϵ_n be any sequence such that $\epsilon_n / \mathrm{Rad}_n(\mathcal{F}) \to \infty$ and $n\epsilon_n \to \infty$. Then, with probability $\to 1$ on the sample, for any $M_n \to \infty$,

$$\hat{\pi}^{\mathcal{F}}_{n,\epsilon_*+\epsilon_n}\left(d_{\mathcal{F}}(P_\theta, P_0) \leq \epsilon_* + \frac{4\epsilon_n}{3} + \mathrm{Rad}_n(\mathcal{F}) + \sqrt{\frac{\log \frac{M_n}{\epsilon_n^d}}{n}} \right) \geq 1 - \frac{2.3^d}{cM_n}.$$

La fin

終わり

ありがとう ございます。