Concentration and robustness of discrepancy-based ABC

Pierre Alquier

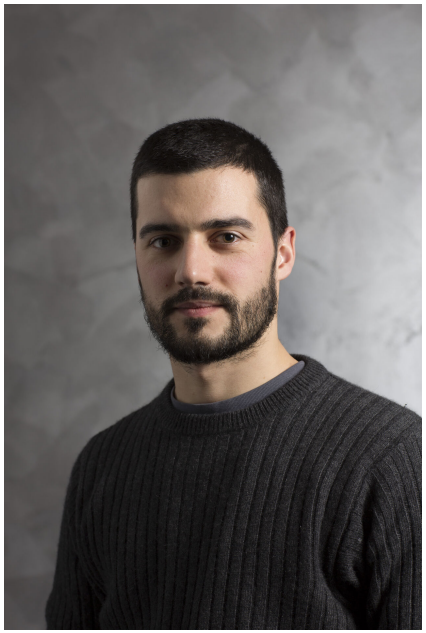
RIKEN AIP
Center for
Advanced Intelligence Project

Séminaire Parisien de Statistique
Institut Henri Poincaré
14 Novembre 2022

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)

Contents

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

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.

Contents

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 P_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 ?

Reminder on IPM

Integral Probability Metrics (IPM)

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

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

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

In general, only a semimetric. However, in many cases, it is actually a metric : $d_{\mathcal{F}}(\mu,\nu) = 0 \Rightarrow \mu = \nu$. Examples :

- total variation : $\mathcal{F} = \{1_A,\ A \text{ measurable}\}$,
- Kolmogorov : $\mathcal{F} = \{1_{(-\infty,x]},\ x \in \mathbb{R}\}$,
- Wasserstein : $\mathcal{F} =$ set of 1-Lipschitz functions,
- Dudley...

Example : 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.

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

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

IPM and statistical estimation

We define the "empirical probability distribution"

$$\hat{\mu}_{Y_{1:n}} := \frac{1}{n} \sum_{i=1}^{n} \delta_{Y_i}.$$

Minimum distance estimator

$$\hat{\theta} := \arg\min_{\theta \in \Theta} d_{\mathcal{F}} \left(\mu_\theta, \hat{\mu}_{Y_{1:n}} \right).$$

Theorem

If $d_{\mathcal{F}}$ is the MMD for a bounded & characteristic kernel,

$$\mathbb{E} \left[d_{\mathcal{F}}(\mu_{\hat{\theta}}, \mu^*) \right] \leq \inf_{\theta \in \Theta} d_{\mathcal{F}}(\mu_\theta, \mu^*) + \frac{2}{\sqrt{n}}.$$

Robust estimation with MMD

$$\mathbb{E}\left[d_{\mathcal{F}}(\mu_{\hat{\theta}}, \mu^*)\right] \leq \inf_{\theta \in \Theta} d_{\mathcal{F}}(\mu_\theta, \mu^*) + \frac{2}{\sqrt{n}}.$$

- well-specified case, $\mu^* = \mu_{\theta_0}$,

$$\mathbb{E}\left[d_{\mathcal{F}}(\mu_{\hat{\theta}}, \mu_{\theta_0})\right] \leq 2/\sqrt{n}.$$

- Huber contamination model $\mu^* = (1-\varepsilon)\mu_{\theta_0} + \varepsilon\nu$,

$$\begin{aligned} d_{\mathcal{F}}(\mu_{\theta_0}, \mu^*) &= \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X \sim \mu_{\theta_0}} f(X) - (1-\varepsilon)\mathbb{E}_{X \sim \mu_{\theta_0}} f(X) - \varepsilon \mathbb{E}_{X \sim \nu} f(X) \right| \\ &= \varepsilon \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X \sim \mu_{\theta_0}} f(X) - \mathbb{E}_{X \sim \nu} f(X) \right| \leq 2\varepsilon \end{aligned}$$

$$\mathbb{E}\left[d_{\mathcal{F}}(\mu_{\hat{\theta}}, \mu_{\theta_0})\right] \leq 4\varepsilon + 2/\sqrt{n}.$$

MDE and robustness : toy experiment

Model : $\mathcal{N}(\theta, 1)$, $X_1, \ldots, X_n$ 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

References on minimum MMD estimation


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


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


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


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


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


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

Contents

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 P_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 $Y_{1:n} = (Y_1, \ldots, Y_n)$ 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

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

(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

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

Reminder on Rademacher complexity

Rademacher complexity

$$\mathfrak{R}_n(\mathcal{F}) := \sup_{\mu} \mathbb{E}_{Y_1,\ldots,Y_n\sim\mu}\, \mathbb{E}_{\varepsilon_1,\ldots,\varepsilon_n} \left[\sup_{f\in\mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(Y_i) \right].$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d Rademacher variables :

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

Examples

- TV : $\mathcal{F} = \{1_A, A \text{ measurable}\}$,

$$\mathfrak{R}_n(\mathcal{F}) \not\to 0 \text{ in general.}$$

- Kolmogorov : $\mathcal{F} = \{1_{(-\infty,x]}, x \in \mathbb{R}\}$,

$$\mathfrak{R}_n(\mathcal{F}) \leq 2\sqrt{\frac{\log(n+1)}{n}} \to 0.$$

- Wasserstein : $\mathcal{F} =$ set of 1-Lipschitz functions,

$$\mathfrak{R}_n(\mathcal{F}) \to 0 \text{ if } \mathcal{X} \text{ is bounded, see Corollary 8 in}$$

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.

- MMD :

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

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 $M_n > 1$,

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

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

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

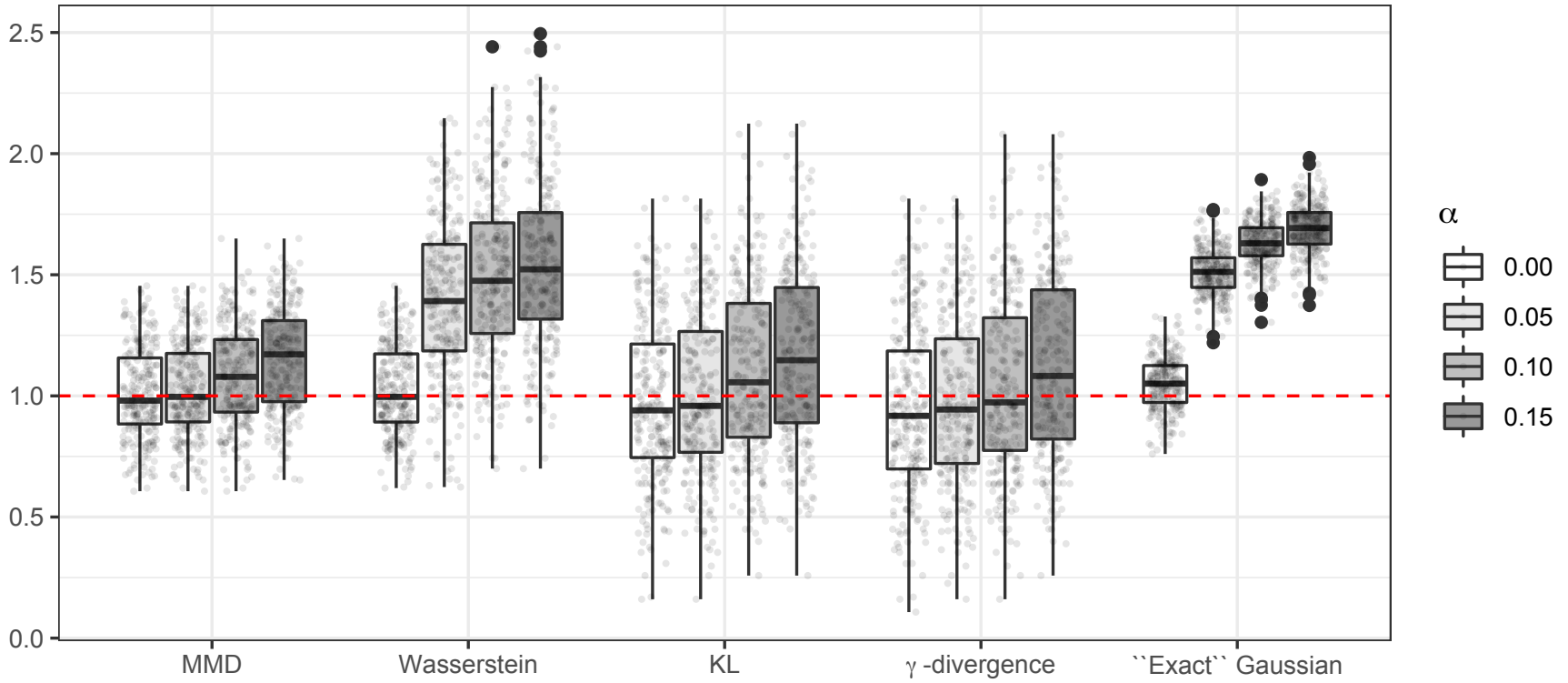
$\epsilon_n = \epsilon^* + \bar{\epsilon}_n$ with $\bar{\epsilon}_n \to 0$. Then, for some $C > 0$, for any sequence $M_n > 1$, with proba. $\to 1$,

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

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



Conclusion

- we provide an analysis of discrepancy-based ABC for a large class of IPM.
- in particular, ABC with MMD leads to robust estimation, without assumptions on the model nor on the truth.
- note that other discrepancies were studied and probably more should be investigated


Frazier, D. T. (2020). Robust and efficient Approximate Bayesian Computation : A minimum distance approach. Preprint arXiv.


Nguyen, H. D., Arbel, J., Lü, H. and Forbes, F. (2020). Approximate Bayesian computation via the energy statistic. IEEE Access.

- important extension to non i.i.d observations (time series, etc.). Note that strong concentration of $d_{\mathcal{F}}\left(\mu_*, \hat{\mu}_{Y_{1:n}}\right)$ is known in this setting (our joint paper with B.-E. Chérief-Abdellatif, Bernoulli 2022).

La fin

終わり

ありがとう ございます。