Minimum Distance Estimation with MMD

Pierre Alquier

ESSEC BUSINESS SCHOOL

UCD Dublin, School of Mathematics and Statistics
Statistics and Actuarial Science seminar
Nov. 30, 2023

1. Minimum Distance Estimation and MMD

2. Applications

3. Algorithms

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

Statistical inference :

- propose a model $(P_\theta, \theta \in \Theta)$, often assume $P^* = P_{\theta^*}$,
- compute $\hat{\theta}_n = \hat{\theta}_n(X_1, \dots, X_n)$.

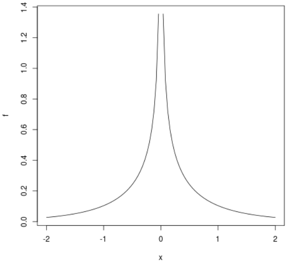
Standard methods :

- Maximum Likelihood Estimators (MLE),
- Bayesian inference,
- method of moments...

Theory of Bayes/MLE strongly relies on $P^* = P_{\theta^*}$, method of moments unstable if the X_i's are heavy-tailed...

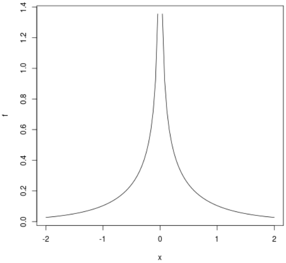
Example :

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

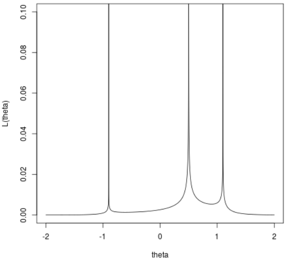


Example :

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



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



What is an outlier ?

Huber proposed the contamination model : with probability ε, X_i is not drawn from P_{θ^*} but from Q that can be anything :

$$P^* = (1-\varepsilon)P_{\theta^*} + \varepsilon Q.$$

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 :

$$P_0 = 0.99 \cdot \mathcal{U}nif[0,1] + 0.01 \cdot \mathcal{U}nif[1000,2000].$$

Minimum Distance Estimation (MDE)

Let $d(\cdot,\cdot)$ be a metric on probability distributions, we put :

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

MDE with well-chosen d leads to robust estimation.

- Wolfowitz, J. (1957). The minimum distance method. The Annals of Mathematical Statistics.
- 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.

- Kolmogorov-Smirnov (KS) suggested by Sture Holm.
- Wasserstein distance studied recently (but not so robust).
- This talk : Maximum Mean Discrepancy (MMD).

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.

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

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

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

Examples

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

Definition : the MMD distance

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

Reminder of the context :

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

Definition - MMD Minimum Distance Estimator

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

Theorem

For any P^*, when $X_1, \dots, X_n$ are i.i.d from P^*,

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

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

- If the model is well-specified, that is $P^* = P_{\theta^*}$,

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

- In Huber contamination model $P^* = (1-\varepsilon)P_{\theta^*} + \varepsilon Q$, using the triangle inequality a few times :

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

Corollary (informally stated)

If the model and the kernel K satisfy, for $\theta' \to \theta$,

$$\mathbb{D}_K\left(P_\theta, P_{\theta'}\right) \sim c \cdot \|\theta - \theta'\|^p$$

then

$$\|\hat{\theta}_K - \theta^*\|^p \sim \frac{2}{c\sqrt{n}} \quad \left(+\frac{4\varepsilon}{c}\right).$$

For example, if $P_\theta = \mathcal{N}(\theta, I_d)$,

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

and thus

$$\|\hat{\theta}_K - \theta^*\|^2 \lesssim \frac{16 d\sigma^2}{n} \quad (+16 d\varepsilon).$$

Setting : we repeat the exp. 200 times,

- $P_\theta = \mathcal{N}(\theta, 1)$,
- $X_1, \dots, X_n$ i.i.d P_{θ^*}, $n = 100$.

	MLE	*MMD*	*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

1. Minimum Distance Estimation and MMD

2. Applications

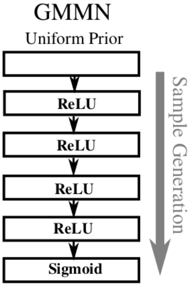
3. Algorithms

1. Minimum Distance Estimation and MMD

2. Applications

3. Algorithms

Application 1 : Generative AI.



Generative model $X \sim P_\theta$:

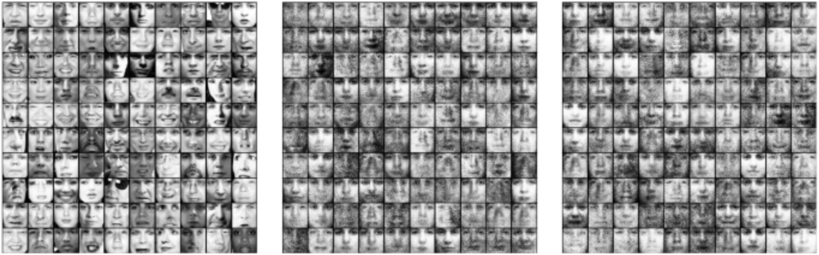
- $U \sim \text{Unif}[0,1]^d$,
- $X = F_\theta(U)$ where F_θ is some neural network with weights θ.

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

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

→ proposed to minimize the MMD to learn θ.

Results from Dziugaite et al. (2015).

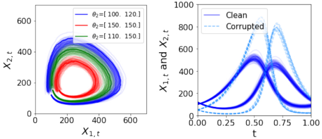


Application 2 : Stochastic Differential Equations.

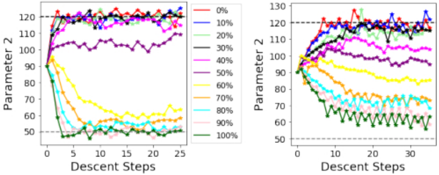
Briol, F. X., Barp, A., Duncan, A. B., & Girolami, M. (2019). Statistical Inference for Generative Models with Maximum Mean Discrepancy. Preprint arXiv :1906.05944.

Stochastic version of Lotka-Volterra model.

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



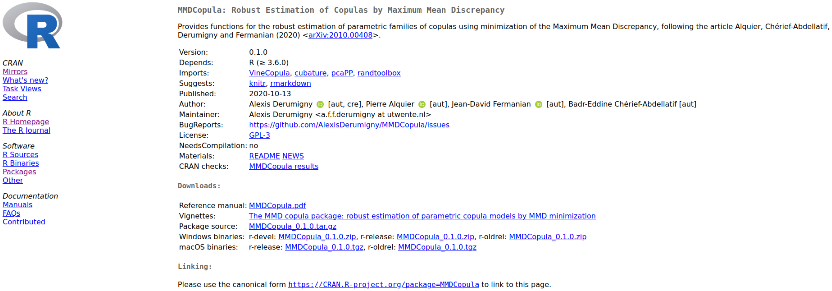
Comparison of MMD // Wasserstein minimization.



Application 3 : Copulas.

$$C_\theta(x_1, x_2) = P_\theta(X_1 \leq F_1^{-1}(x_1), X_2 \leq F_2^{-1}(x_2)).$$

Alquier, P., Chérief-Abdellatif, B.-E., Derumigny, A. and Fermanian, J.-D. (2023). Estimation of copulas via Maximum Mean Discrepancy. JASA.



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>.

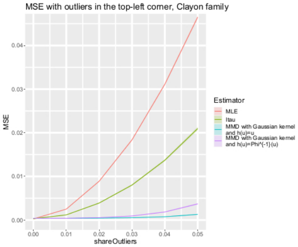
Version:	0.1.0
Depends:	R (≥ 3.6.0)
Imports:	VineCopula, cubature, pcaPP, randtoolbox
Suggests:	knitr, rmarkdown
Published:	2020-10-13
Author:	Alexis Derumigny [aut, cre], Pierre Alquier [aut], Jean-David Fermanian [aut], Badr-Eddine Chérief-Abdellatif [aut]
Maintainer:	Alexis Derumigny <a.f.f.derumigny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

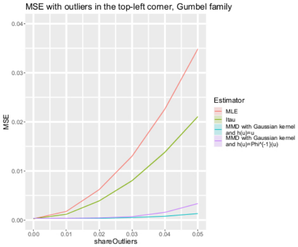
`Downloads:`

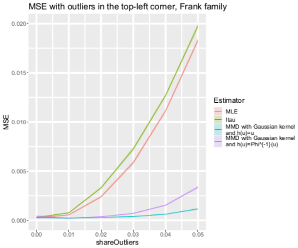
Reference manual:	MMDCopula.pdf
Vignettes:	The MMD copula package: robust estimation of parametric copula models by MMD minimization
Package source:	MMDCopula_0.1.0.tar.gz
Windows binaries:	r-devel: MMDCopula_0.1.0.zip, r-release: MMDCopula_0.1.0.zip, r-oldrel: MMDCopula_0.1.0.zip
macOS binaries:	r-release: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


MSE with outliers in the top-left corner, Clayon family


MSE with outliers in the top-left corner, Gumbel family

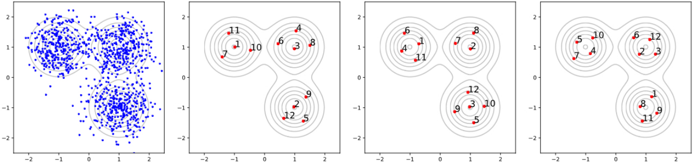

MSE with outliers in the top-left corner, Frank family

Application 4 : quantization / data compression.

Teymur, O., Gorham, J., Riabiz, M. and Oates, C. (2021). Optimal quantisation of probability measures using maximum mean discrepancy. AISTAT.

Idea : compress a sample $X_1, \dots, X_n$ to $x_1, \dots, x_m$ with $m \ll n$, by

$$\min_{\theta=(x_1,\dots,x_m)} \mathbb{D}_K\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}, \frac{1}{m}\sum_{j=1}^{m} \delta_{x_j}\right).$$



Application/extension 5 : time series.

Theorem

When $X_1, \dots, X_n$ are ~~i.i.d~~ from a stationary time series with marginal distribution P^*,

$$\mathbb{E}\left[\mathbb{D}_K\left(P_{\hat{\theta}_K}, P^*\right)\right] \leq \inf_{\theta\in\Theta} \mathbb{D}_K(P_\theta, P^*) + \frac{2\sqrt{1+2\sum_{s>0}\rho_s}}{\sqrt{n}}$$

where

$$\rho_s = \left|\mathbb{E}K(X_0, X_s) - \mathbb{E}K(X_0, X_0')\right|.$$

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

For classical processes (Markov chains and hidden Markov chains, autoregressions etc), $\sum_{s>0}\rho_s < +\infty$.

Application/extension 6 : regression.

- Assume we observe $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- A direct application of the above allows to estimate the joint distribution of (X, Y).
- But we are interested in $Y|X$, and we don't want to specify a model for X !
- Feasible, but far more complicated as the theory of conditional mean embeddings is difficult...

Alquier, P. and Gerber, M. (2024). Universal Robust Regression via Maximum Mean Discrepancy. To appear in Biometrika.

1 Minimum Distance Estimation and MMD

2 Applications

3 Algorithms

Reminder

$$\mu_K(P) = \mathbb{E}_{X\sim P}[\Phi(X)] \qquad K(u,v) = \langle \Phi(u), \Phi(v) \rangle_{\mathcal{H}} \, ,$$

$$\mathbb{D}_K^2(P,Q) = \left\| \mu_K(P) - \mu_K(Q) \right\|_{\mathcal{H}}^2 .$$

$$\mathbb{D}_K^2(P,Q) = \left\| \mu_K(P) \right\|_{\mathcal{H}}^2 - 2 \left\langle \mu_K(P), \mu_K(Q) \right\rangle_{\mathcal{H}} + \left\| \mu_K(Q) \right\|_{\mathcal{H}}^2 .$$

For example,

$$\begin{aligned} \left\langle \mu_K(P), \mu_K(Q) \right\rangle_{\mathcal{H}} & = \left\langle \mathbb{E}_{X\sim P}[\Phi(X)], \mathbb{E}_{X'\sim Q}[\Phi(X')] \right\rangle_{\mathcal{H}} \\ & = \mathbb{E}_{X\sim P, X'\sim Q} \left\langle \Phi(X), \Phi(X') \right\rangle_{\mathcal{H}} \\ & = \mathbb{E}_{X\sim P, X'\sim Q} K(X, X'). \end{aligned}$$

$$\begin{aligned} \mathbb{D}_K^2(P,Q) = \mathbb{E}_{X\sim P, Y\sim P} K(X,Y) - 2 \mathbb{E}_{X\sim P, X'\sim Q} K(X,X') \\ + \mathbb{E}_{X'\sim Q, Y'\sim Q} K(X,X'). \end{aligned}$$

Reminder

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

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

- Exact evaluation sometimes possible if we have a close formula for $\mathbb{E}_{X\sim P_\theta}$...
- Monte-Carlo approx. only requires to sample $X \sim P_\theta$.

Theorem : gradient of MMD

Assume P_θ has a differentiable density p_θ, then

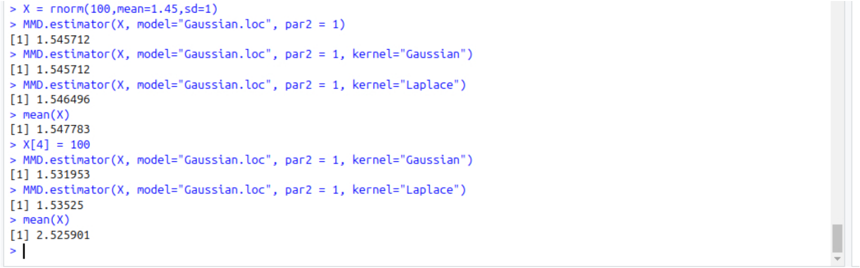
$$\nabla_\theta \mathbb{D}_K^2(P_\theta, \hat{P}_n) = 2\mathbb{E}_{X,X'\sim P_\theta}\left\{\left(K(X,X') - \frac{1}{n}\sum_{i=1}^n K(X_i,X)\right)\nabla_\theta \log p_\theta(X)\right\}.$$

If we sample $X, X' \sim P_\theta$, then

$$\hat{\nabla} := \left(K(X,X') - \frac{1}{n}\sum_{i=1}^n K(X_i,X)\right)\nabla_\theta \log p_\theta(X)$$

is an unbiased estimator of $\nabla_\theta \mathbb{D}_K^2(P_\theta, \hat{P}_n)$. We can thus minimize by stochastic gradient.

R package (joint work with Mathieu Gerber) available soon. Will contain stochastic optimization to compute $\hat{\theta}_K$ in 15-20 classical parametric models + regression, logistic regression, etc.



```
> X = rnorm(100,mean=1.45,sd=1)
> MMD.estimator(X, model="Gaussian.loc", par2 = 1)
[1] 1.545712
> MMD.estimator(X, model="Gaussian.loc", par2 = 1, kernel="Gaussian")
[1] 1.545712
> MMD.estimator(X, model="Gaussian.loc", par2 = 1, kernel="Laplace")
[1] 1.546496
> mean(X)
[1] 1.547783
> X[4] = 100
> MMD.estimator(X, model="Gaussian.loc", par2 = 1, kernel="Gaussian")
[1] 1.531953
> MMD.estimator(X, model="Gaussian.loc", par2 = 1, kernel="Laplace")
[1] 1.53525
> mean(X)
[1] 2.525901
> |
```

"Bayesian style" estimators (and algorithms) based on MMD ?

1) MMD-Bayes

$$\pi(\theta|X_1,\dots,X_n) \propto \pi(\theta) \exp\left[-\beta \cdot \mathbb{D}_K\left(P_\theta, \hat{P}_n\right)\right]$$

- $\pi(\theta|X_1,\dots,X_n)$ approximated by variational inference (using stochastic gradient again)

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

- pseudo-marginal MCMC / noisy MCMC :

Pacchiardi, L. and Dutta, R. (2021). Generalized Bayesian likelihood-free inference using scoring rules estimators. ArXiv preprint arXiv :2104.03889.

2) MMD-ABC

```
INPUT : data $X_1, \dots, X_n$, model $(P_\theta, \theta \in \Theta)$, prior $\pi$ and threshold $\epsilon$.
  (i) sample $\theta \sim \pi$,
  (ii) sample $Z_1, \dots, Z_n$ i.i.d. from $P_\theta$,
      • if $\mathbb{D}_K(\frac{1}{n}\sum_{i=1}^n \delta_{Z_i}, \frac{1}{n}\sum_{i=1}^n \delta_{X_i}) \leq \epsilon$ return $\theta$,
      • else goto (i).
OUTPUT : $\vartheta$.
```

- The distribution $\pi_{\text{MMD-ABC}}(\vartheta)$ of the output ϑ converges to the (usual) posterior when $\epsilon \to 0$,
- $\epsilon > 0$ computationally more efficient AND robust...

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

Open questions :

- faster algorithms, especially when likelihood not available,
- general results to connect $\|\theta_1 - \theta_2\|$ to $\mathbb{D}_K(P_{\theta_1}, P_{\theta_2})$ in non-standard models,
- data-driven kernel choice ? Does it always lead to optimal rates ?

Much tighter bounds on $\mathbb{D}_K\left(P_{\hat{\theta}_K}, P^*\right)$ in the following paper. Helps to understand the role of the kernel K :

Wolfer, G. and Alquier, P. (2022). Variance-Aware Estimation of Kernel Mean Embedding. Preprint arXiv :2210.06672.

Thank you !