Approximate Bayesian computation via the energy statistic

Hien D. Nguyen[1]

[1]Senior Lecturer, La Trobe University, Melbourne Australia.
(Email: h.nguyen5@latrobe.edu.au; Website: hiendn.github.io)

One World ABC Seminar, 18 June, 2020

LA TROBE
UNIVERSITY

Acknowledgements

From left to right: Julyan Arbel, Florence Forbes, Hongliang Lü

(STATIFY Team, Inria Grenoble Rhône-Alpes)

Outline

- A *very brief* introduction to approximate Bayesian computation
- The energy distance and its estimator: **the energy statistic**
- The **energy statistic ABC** and its properties
- Numerical illustrations and comparisons to other methods

Preliminary setup

- Let $\mathbf{X}_n = \{\boldsymbol{X}_i\}_{i=1}^n$ be an IID sample of n replicates of the random variable $\boldsymbol{X} \in \mathbb{X} \subseteq \mathbb{R}^q$, where:
  - The DGP of $\boldsymbol{X}$ has parametric PDF $f(\boldsymbol{x}|\boldsymbol{\theta})$.
  - The parameter $\boldsymbol{\theta} \in \mathbb{T}$ has prior PDF $\pi(\boldsymbol{\theta})$.
- Conditional on $\boldsymbol{\theta}$, the likelihood of $\mathbf{X}_n$ is

$$f(\mathbf{x}_n|\boldsymbol{\theta}) = \prod_{i=1}^n f(\boldsymbol{x}_i|\boldsymbol{\theta}).$$

- We aim to compute the PDF of the posterior distribution:

$$\pi(\boldsymbol{\theta}|\mathbf{x}_n) = c^{-1}(\mathbf{x}_n) f(\mathbf{x}_n|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}),$$

  where $\mathbf{x}_n$ is a realization of $\mathbf{X}_n$, and

$$c(\mathbf{x}_n) = \int_{\mathbb{T}} f(\mathbf{x}_n|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

Simulation-based inference

- Suppose that $f(\boldsymbol{x}|\boldsymbol{\theta})$ is infeasible to compute, but it is feasible to simulate size m IID samples $\mathbf{Y}_m = \{\boldsymbol{Y}_i\}_{i=1}^m$ with PDF

$$f(\mathbf{y}_n|\boldsymbol{\theta}) = \prod_{i=1}^{m} f(\boldsymbol{y}_i|\boldsymbol{\theta}).$$

- Let $\boldsymbol{\theta}_k$ be simulated from PDF $\pi(\boldsymbol{\theta})$, and let $\mathbf{Y}_{m,k}$ be the corresponding sample simulated from PDF $f(\mathbf{y}_n|\boldsymbol{\theta}_k)$ and generate N pairs $\boldsymbol{Z}_{m,k} = (\mathbf{Y}_{m,k}, \boldsymbol{\theta}_k)$:

$$\mathbf{Z}_N = \left\{\boldsymbol{Z}_{m,k}\right\}_{k=1}^{N}.$$

- We wish to approximate the posterior $\pi(\boldsymbol{\theta}|\mathbf{x}_n)$ using $\mathbf{Z}_N$.

Approximating the posterior

- Let $\mathscr{D}(\mathbf{x}_n, \mathbf{y}_m) \geq 0$ be a *discrepancy function* that measures the difference between the distributions of $\mathbf{x}_n$ and $\mathbf{y}_m$.
- Let $w(d, \varepsilon) \geq 0$ be a *weight function*, decreasing in $d \geq 0$ and calibrated by $\varepsilon > 0$.
- Approximate the likelihood $f(\mathbf{x}_n|\boldsymbol{\theta})$ by

$$L_{m,\varepsilon}(\mathbf{x}_n|\boldsymbol{\theta}) = \int_{\mathbb{X}^m} w(\mathscr{D}(\mathbf{x}_n, \mathbf{y}_m), \varepsilon) f(\mathbf{y}_m|\boldsymbol{\theta}) \mathrm{d}\mathbf{y}_m.$$

- In the language of Jiang et al. (2018), we have the **pseudo-posterior** PDF approximation

$$\pi_{m,\varepsilon}(\boldsymbol{\theta}|\mathbf{x}_n) = c_{m,\varepsilon}^{-1}(\mathbf{x}_n) \pi(\boldsymbol{\theta}) L_{m,\varepsilon}(\mathbf{x}_n|\boldsymbol{\theta}),$$

where $c_{m,\varepsilon}(\mathbf{x}_n) = \int_{\mathbb{T}} \pi(\boldsymbol{\theta}) L_{m,\varepsilon}(\mathbf{x}_n|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$.

The ABC algorithm

Input: Data $\mathbf{x}_n$; Discrepancy function $\mathscr{D}$; Weight function w; Calibration parameter ε; Sample size m.

For $k \in \{1, \ldots, N\}$;

Simulate $\boldsymbol{\theta}_k$ from PDF $\pi(\boldsymbol{\theta})$;

Simulate $\boldsymbol{Y}_{m,k}$ from PDF $f(\mathbf{y}_m|\boldsymbol{\theta}_k)$;

Compute discrepancy $D_k = \mathscr{D}(\mathbf{x}_n, \mathbf{Y}_{m,k})$;

Put $\boldsymbol{Z}_{m,k} = (\mathbf{Y}_{m,k}, \boldsymbol{\theta}_k)$ into $\mathbf{Z}_N$ and D_k into $\mathbf{D}_N = \{D_k\}_{k=1}^N$.

Output: $\mathbf{Z}_N$ and $\mathbf{D}_N$; Approximate the pseudo-posterior distribution by the discrete measure

$$\Pi_{m,\varepsilon}^N(\boldsymbol{\theta}|\mathbf{x}_n) = \left[\sum_{k=1}^N w(D_k, \varepsilon)\right]^{-1} \sum_{k=1}^N w(D_k, \varepsilon)\mathbb{I}\{\boldsymbol{\theta} = \boldsymbol{\theta}_k\}.$$

The energy distance

- For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{X}$, let $\delta(\boldsymbol{x}, \boldsymbol{y})$ be a **semi-metric of negative type**, in the sense of Sejdinovic et al. (2013).
- For independent $\boldsymbol{X}, \boldsymbol{X}'$ and $\boldsymbol{Y}, \boldsymbol{Y}'$ with probability measures Π_X and Π_Y, respectively, the **energy distance** (ED) based on δ is $\mathscr{E}_\delta^{1/2}$, where:

$$\mathscr{E}_\delta(\Pi_X, \Pi_Y) = 2\mathbb{E}[\delta(\boldsymbol{X}, \boldsymbol{Y})] - \mathbb{E}\left[\delta\left(\boldsymbol{X}, \boldsymbol{X}'\right)\right] - \mathbb{E}\left[\delta\left(\boldsymbol{Y}, \boldsymbol{Y}'\right)\right].$$

- Szekely and Rizzo (2017) states that the ED satisfies:
  1. $\mathscr{E}_\delta(\Pi_X, \Pi_Y) \geq 0$,
  2. $\mathscr{E}_\delta(\Pi_X, \Pi_Y) = 0 \iff \Pi_X = \Pi_Y$,

  under the condition that

$$\mathbb{E}\left[\delta\left(\boldsymbol{X}, \boldsymbol{X}'\right)\right] + \mathbb{E}\left[\delta\left(\boldsymbol{Y}, \boldsymbol{Y}'\right)\right] < \infty.$$

The Euclidean case

- The original paper of Szekely and Rizzo (2004) studied the $\delta(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$ case, and we denote the squared ED, in this case, by $\mathscr{E}$.
- Let φ_X and φ_Y be the characteristic functions of Π_X and Π_Y, respectively. Then, under the condition that $\mathbb{E}\|\mathbf{X}\| + \mathbb{E}\|\mathbf{Y}\| < \infty$, $\mathscr{E}$ has the closed form:

$$\mathscr{E}(\Pi_X, \Pi_Y) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{(d+1)/2}} \int_{\mathbb{R}^d} \frac{|\varphi_X(\boldsymbol{t}) - \varphi_Y(\boldsymbol{t})|^2}{\|\boldsymbol{t}\|_2^{d+1}} \mathrm{d}\boldsymbol{t}.$$

- Via the characterization, $\mathscr{E}^{1/2}$ *is a metric* over set space of probability measures with finite first moment.

The energy statistic

- Given samples $\mathbf{X}_n$ and $\mathbf{Y}_m$, from measures Π_X and Π_Y, one can estimate $\mathscr{E}_\delta(\Pi_X, \Pi_Y)$ by the energy V-statistic (ES):

$$\begin{aligned}\mathscr{V}_\delta(\mathbf{X}_n, \mathbf{Y}_m) &= \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\delta(\boldsymbol{X}_i, \boldsymbol{Y}_j) \\ &\quad -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta(\boldsymbol{X}_i, \boldsymbol{X}_j) \\ &\quad -\frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\delta(\boldsymbol{Y}_i, \boldsymbol{Y}_j).\end{aligned}$$

- For brevity, we write $\mathscr{V}_\delta = \mathscr{V}$ when δ is the Euclidean norm, and note that $\mathscr{V}^{1/2}$ is a metric over the set of finite discrete measures (Szekely and Rizzo, 2017).
- We call the use of $\mathscr{D}(\mathbf{X}_n, \mathbf{Y}_m) = \mathscr{V}(\mathbf{X}_n, \mathbf{Y}_m)$ the **ES ABC algorithm**.

Relationship to the MMD ABC

- In Park et al. (2016), Jiang et al. (2018), and Bernton et al. (2019), the maximum mean discrepancy (MMD) ABC algorithms were considered, where one uses

$$\begin{aligned}\mathscr{D}(\mathbf{X}_n, \mathbf{Y}_m) &= -\frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\chi(\boldsymbol{X}_i, \boldsymbol{Y}_j) \\ &\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\chi(\boldsymbol{X}_i, \boldsymbol{X}_j) \\ &\quad + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}\chi(\boldsymbol{Y}_i, \boldsymbol{Y}_j),\end{aligned}$$

  for some Mercer kernel χ.
- Sejdinovic et al. (2013) established that Mercer kernels generative relationship between Mercer kernels and semi-metrics of negative type.

A general asymptotic result

Let $\mathbf{X}_n$ and $\mathbf{Y}_m$ be IID samples with PDFs $f(\mathbf{x}_n|\boldsymbol{\theta}_0)$ and $f(\mathbf{y}_m|\boldsymbol{\theta})$. Assume that $\mathscr{D}(\mathbf{X}_n, \mathbf{Y}_m)$ converges to $\mathscr{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, almost surely as $n, m(n) \to \infty$. If $w(d, \varepsilon)$ is piecewise continuous and bounded for all $d, \varepsilon > 0$, and if $w(\cdot, \varepsilon)$ is continuous at $\mathscr{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$, then

$$\pi_{m(n),\varepsilon}(\boldsymbol{\theta}|\mathbf{X}_n) \to \frac{\pi(\boldsymbol{\theta})\, w(\mathscr{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \varepsilon)}{\int_{\mathbb{T}} \pi(\boldsymbol{\theta})\, w(\mathscr{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta}), \varepsilon)\, \mathrm{d}\boldsymbol{\theta}},$$

almost surely, as $n \to \infty$.

- Jiang et al. (2018) proved the rejection case:

$$w(d, \varepsilon) = \mathbb{I}\{d < \varepsilon\}.$$

Verification of the ES case

- A sufficient condition to verify the almost sure convergence of $\mathscr{D}(\mathbf{X}_n, \mathbf{Y}_m)$ to $\mathscr{D}_\infty(\boldsymbol{\theta}_0, \boldsymbol{\theta})$ is for Π_X and Π_Y to have finite second moments:

$$\mathbb{E}\left(\|\boldsymbol{X}\|_2^2\right) + \mathbb{E}\left(\|\boldsymbol{Y}\|_2^2\right) < \infty.$$

- In this case, we have:

$$\mathscr{D}(\mathbf{X}_n, \mathbf{Y}_m) \to \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{(d+1)/2}} \int_{\mathbb{R}^d} \frac{|\varphi(\boldsymbol{t}; \boldsymbol{\theta}_0) - \varphi(\boldsymbol{t}; \boldsymbol{\theta})|^2}{\|\boldsymbol{t}\|_2^{d+1}} \mathrm{d}\boldsymbol{t},$$

almost surely, as $n \to \infty$, where Π_X and Π_Y are characterized by $\varphi(\boldsymbol{t}; \boldsymbol{\theta}_0)$ and $\varphi(\boldsymbol{t}; \boldsymbol{\theta})$.

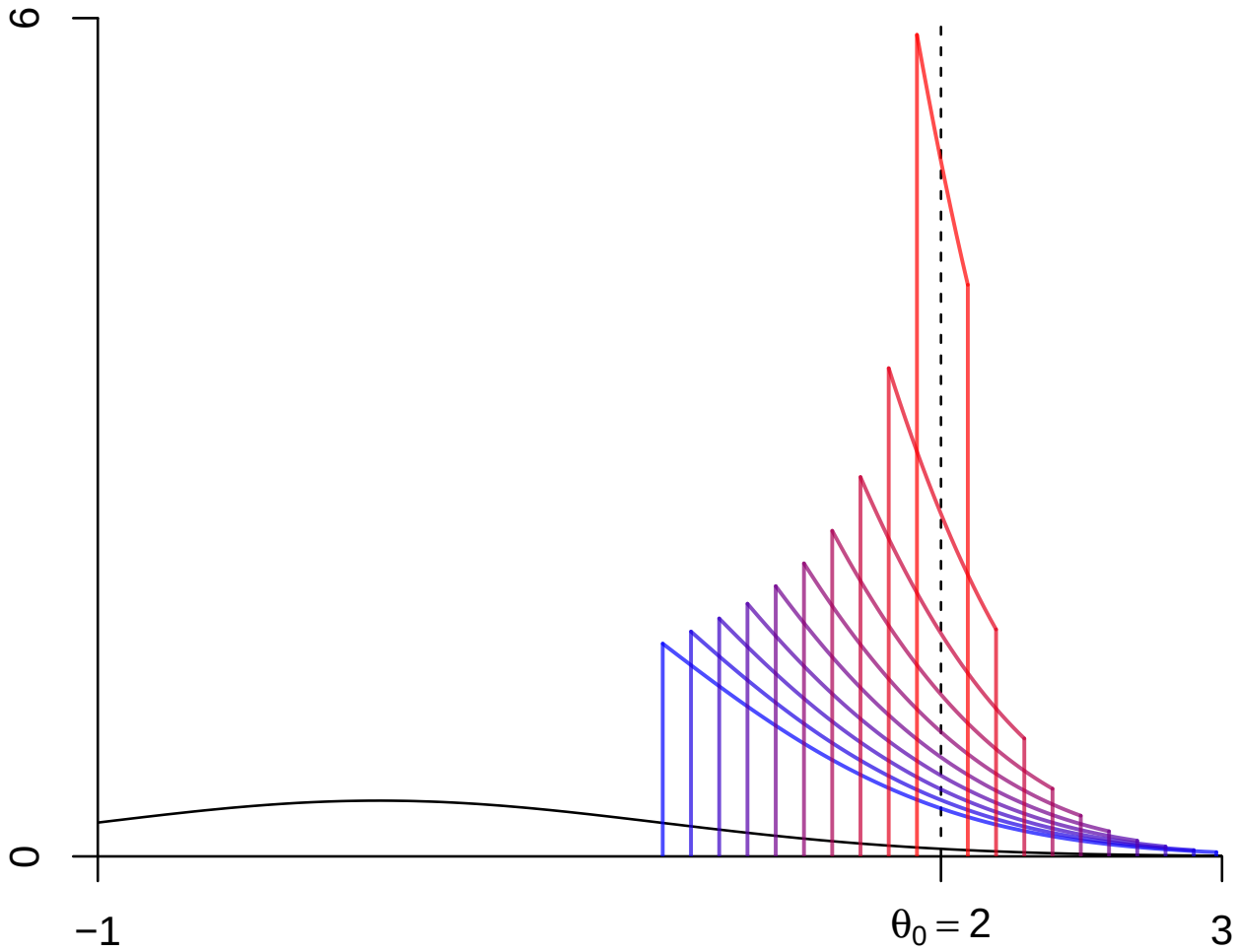
A toy example

- Suppose that $\theta \sim \mathrm{N}\left(0, \tau^2\right)$
- We observe a sample $\mathbf{X}_n$ of IID replications of $X|\theta_0 \sim \mathrm{N}\left(\theta_0, \sigma^2\right)$, with $\theta_0 = 2$.
- The analytic solution for the posterior is that $\theta|\mathbf{X}_n \sim \mathrm{N}\left(\hat{\theta}, \hat{\sigma}^2\right)$, where

$$\hat{\theta} = \frac{n\bar{X}_n}{n + \sigma^2/\tau^2}, \; \hat{\sigma}^{-2} = \frac{1}{\tau^2} + \frac{n}{\sigma^2}.$$

- Note that $\mathscr{D}_\infty(\theta_0, \theta) = (\theta_0 - \theta)^2$ in the ES case.
- We consider the ES ABC pseudo-posterior, when we use simulated samples $\mathbf{Y}_m$ of IID replications of $Y|\theta \sim \mathrm{N}\left(\theta, \sigma^2\right)$.

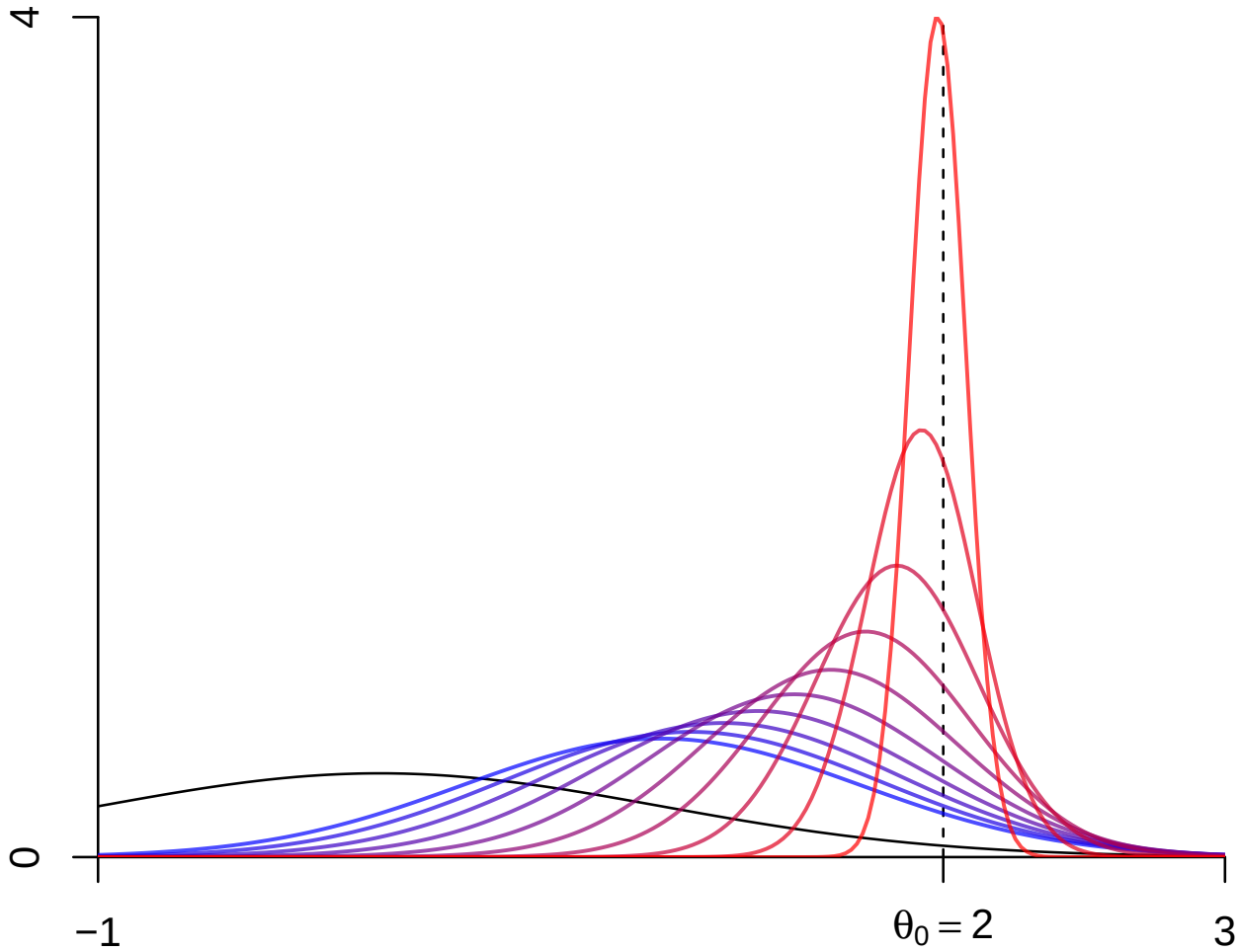
A toy example



ES ABC pseudo-posterior limit using the rejection weights:

$$w(d, \varepsilon) = \mathbb{I}\{d < \varepsilon\}.$$

A toy example



ES ABC pseudo-posterior limit using the Gaussian weights:

$$w(d, \varepsilon) = \exp\left(-d^2/2\varepsilon^2\right).$$

A finite sample result

Let $n = m$, and $w(d, \varepsilon) = \mathbb{I}\{d < \varepsilon\}$. Assume that $f(\mathbf{x}_n|\boldsymbol{\theta})$ is continuous and exchangeable, and that

$$\sup_{\boldsymbol{\theta} \in \mathbb{T} \setminus \{\Theta \subset \mathbb{T} : \int_\Theta \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 0\}} f(\mathbf{x}_n|\boldsymbol{\theta}) < \infty,$$

and

$$\sup_{\boldsymbol{\theta} \in \mathbb{T} \setminus \{\Theta \subset \mathbb{T} : \int_\Theta \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} = 0\}} \sup_{\{\mathbf{y}_n : \mathscr{D}(\mathbf{x}_n, \mathbf{y}_n) \leq \bar{\varepsilon}\}} f(\mathbf{y}_n|\boldsymbol{\theta}) < \infty,$$

for some $\bar{\varepsilon} > 0$. Then, for fixed $\mathbf{x}_n$, the ES ABC pseudo-posterior PDF converges strongly to $\pi(\boldsymbol{\theta}|\mathbf{x}_n)$, as $\varepsilon \to 0$.

- This result along with others regarding convergence when $\varepsilon \to 0$ and $n \to \infty$, simultaneously, can be taken directly from Bernton et al. (2019), due to the metric property of $\mathscr{V}^{1/2}$.

Numerical setup

Set $w(d, \varepsilon) = \mathbb{I}\{d < \varepsilon\}$ and $\mathscr{D} = \mathscr{V}$;

Set $N = 10^5$;

Set $m = n$;

Set $\varepsilon = \mathscr{Q}_N(0.05)$, where $\mathscr{Q}_N$ is the quantile empirical quantile function of $\mathbf{D}_N$.

For comparison, we also consider the following alternative for $\mathscr{D}$:

- the MMD discrepancy, with Gaussian kernel χ (Park et al., 2016),
- the Kullback–Leibler (KL) discrepancy (Jiang et al., 2018),
- the "swapping distance", approximation to the $\mathscr{L}_2$ Wasserstein (WA) discrepancy (Bernton et al., 2019).

Numerical study: Gaussian mixture

- We observe realization $\mathbf{x}_n$ of $\mathbf{X}_n$, $n = 500$, containing observations

$$\boldsymbol{X}_i \sim p^* \mathrm{N}(\boldsymbol{\mu}_0^*, \boldsymbol{\Sigma}_0) + (1 - p^*) \mathrm{N}(\boldsymbol{\mu}_1^*, \boldsymbol{\Sigma}_1),$$

where $p^* = 0.3$, $\boldsymbol{\mu}_0^* = (0.7, 0.7)$, $\boldsymbol{\mu}_1^* = -\boldsymbol{\mu}_0^*$,

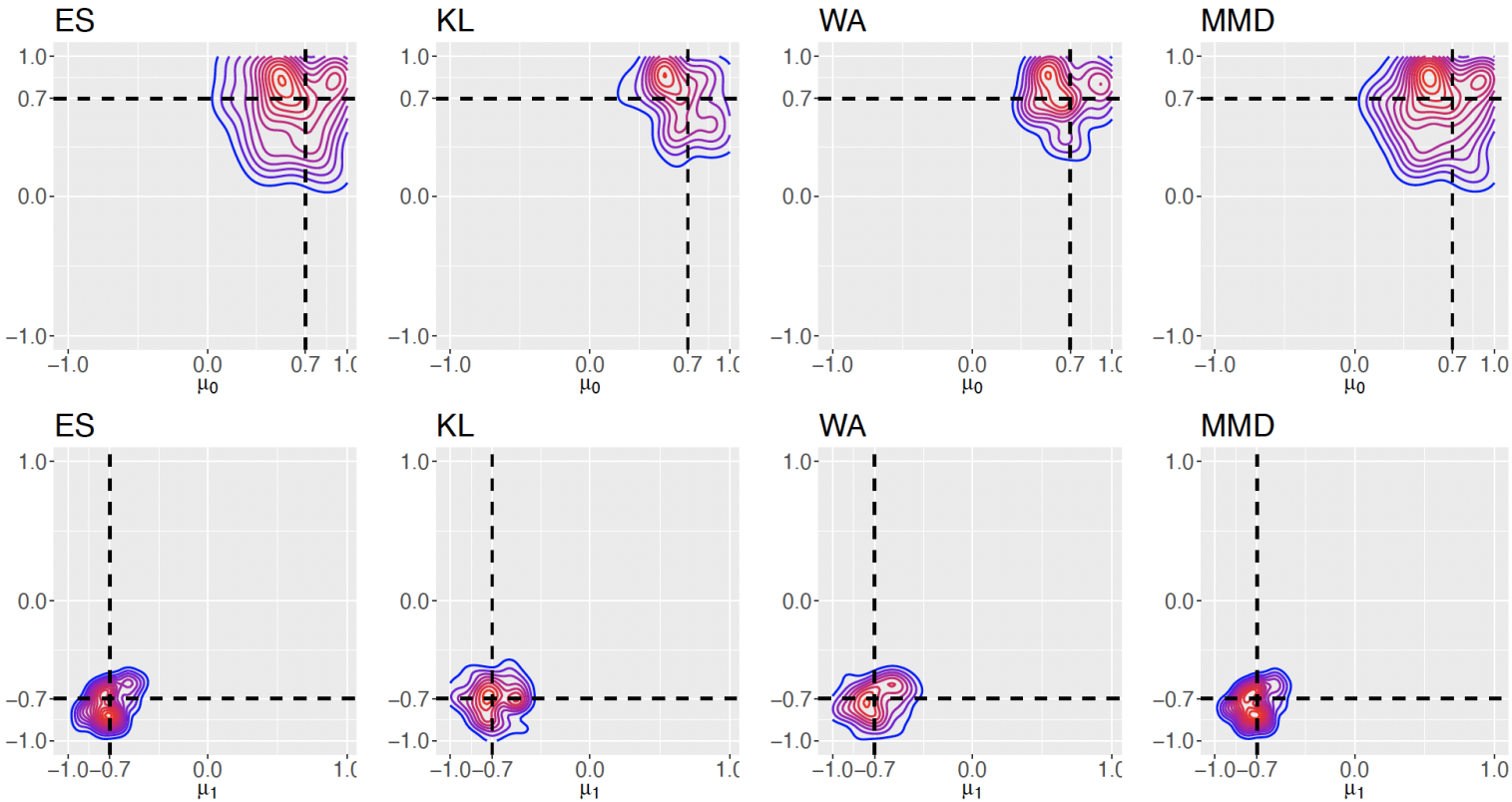
$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{bmatrix}, \text{ and } \boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix}.$$

- We wish to approximate $\pi(\boldsymbol{\theta}|\mathbf{x}_n)$, where $\boldsymbol{\theta} = (p, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1)$.

Simulate $p_k \sim \mathrm{U}(0,1)$; $\boldsymbol{\mu}_{0,k}, \boldsymbol{\mu}_{1,k} \sim \mathrm{U}(-1,1)^2$;

Simulate $\boldsymbol{Y}_{n,k} \sim p_k \mathrm{N}(\boldsymbol{\mu}_{0,k}, \boldsymbol{\Sigma}_0) + (1 - p_k) \mathrm{N}(\boldsymbol{\mu}_{1,k}, \boldsymbol{\Sigma}_1)$.

Numerical study: Gaussian mixture



Kernel density estimates of the ABC-obtained posterior samples.

Numerical study: moving average

- We observe realization $\mathbf{x}_n$ of $\mathbf{X}_n$, $n = 200$, containing observations $\boldsymbol{X}_i \in \mathbb{R}^{10}$, such that

$$X_{i,t} = Z_t + \sum_{j=1}^{2} \theta_j^* Z_{t-j}, \tag{1}$$

  where $\{Z_t\}_{t\in\mathbb{Z}}$ is an IID sequence of Student-t noise, with 5 degrees of freedom, and

$$\boldsymbol{\theta}^* = (\theta_1^*, \theta_2^*) = (0.6, 0.2)\,.$$

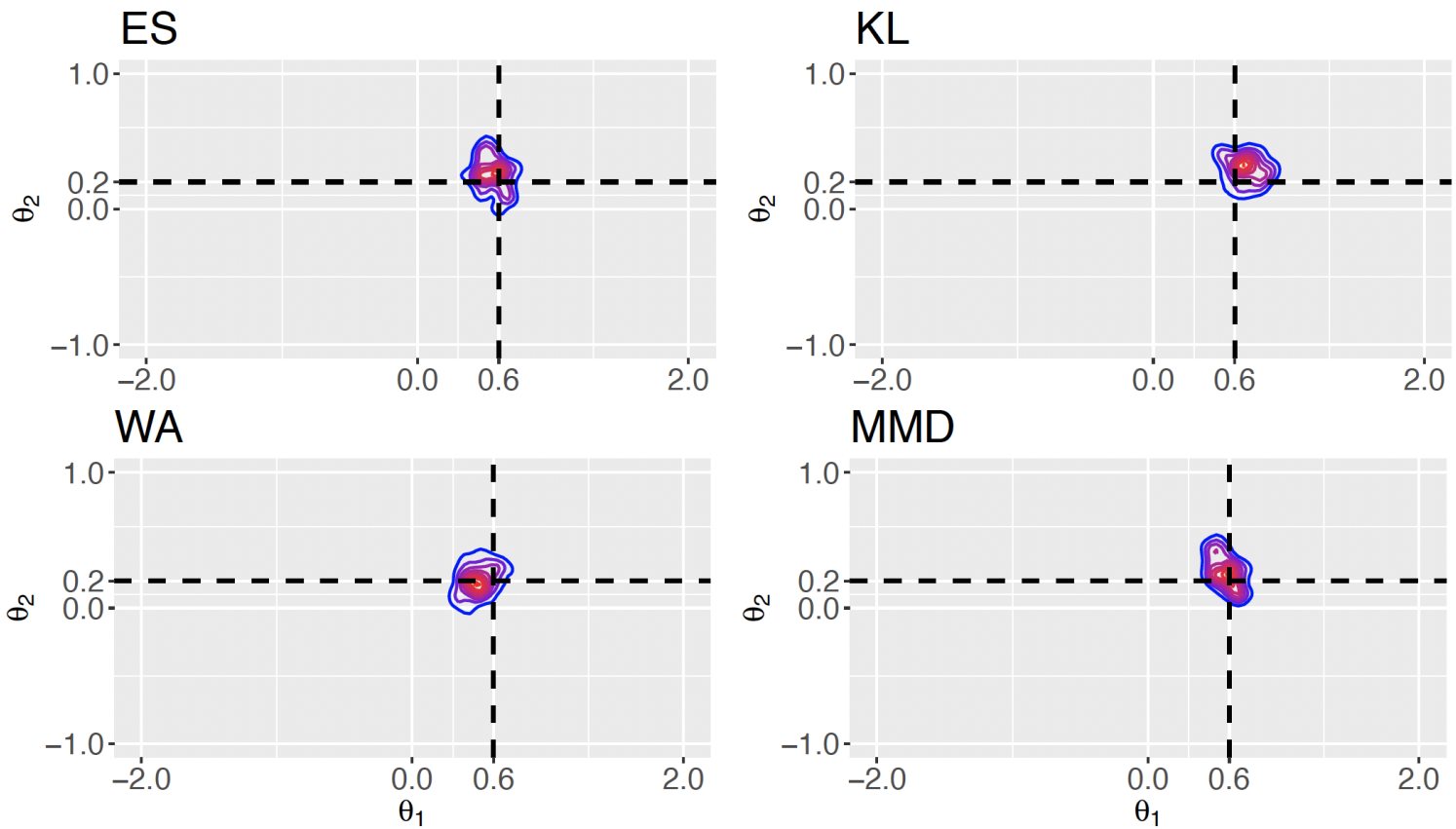
- We wish to approximate $\pi(\boldsymbol{\theta}|\mathbf{x}_n)$.

Simulate $\theta_{1,k} \sim \mathrm{U}(-2,2)$; $\theta_{2,k} \sim \mathrm{U}(-1,1)$;

Simulate $\boldsymbol{Y}_{n,k} \in \mathbb{R}^{10}$ from (1), with $\boldsymbol{\theta}^*$ replaced by

$$\boldsymbol{\theta}_k = (\theta_{1,k}, \theta_{2,k})\,.$$

Numerical study: moving average



Kernel density estimates of the ABC-obtained posterior samples.

Numerical study: bivariate beta distribution

- We observe realization $\mathbf{x}_n$ of $\mathbf{X}_n$, $n = 500$, containing observations $\boldsymbol{X}_i \in \mathbb{R}^2$, such that

$$X_{i,1} \sim \text{Beta}\left(\theta_1^* + \theta_3^*, \theta_4^* + \theta_5^*\right),\ X_{i,2} \sim \text{Beta}\left(\theta_2^* + \theta_4^*, \theta_3^* + \theta_5^*\right), \tag{2}$$

  where $\boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_5^*) = \mathbf{1}$.

- We wish to approximate $\pi(\boldsymbol{\theta}|\mathbf{x}_n)$.

Simulate $\boldsymbol{\theta}_k = (\theta_{k,1}, \ldots, \theta_{k,5}) \sim \text{U}(0,5)^5$;

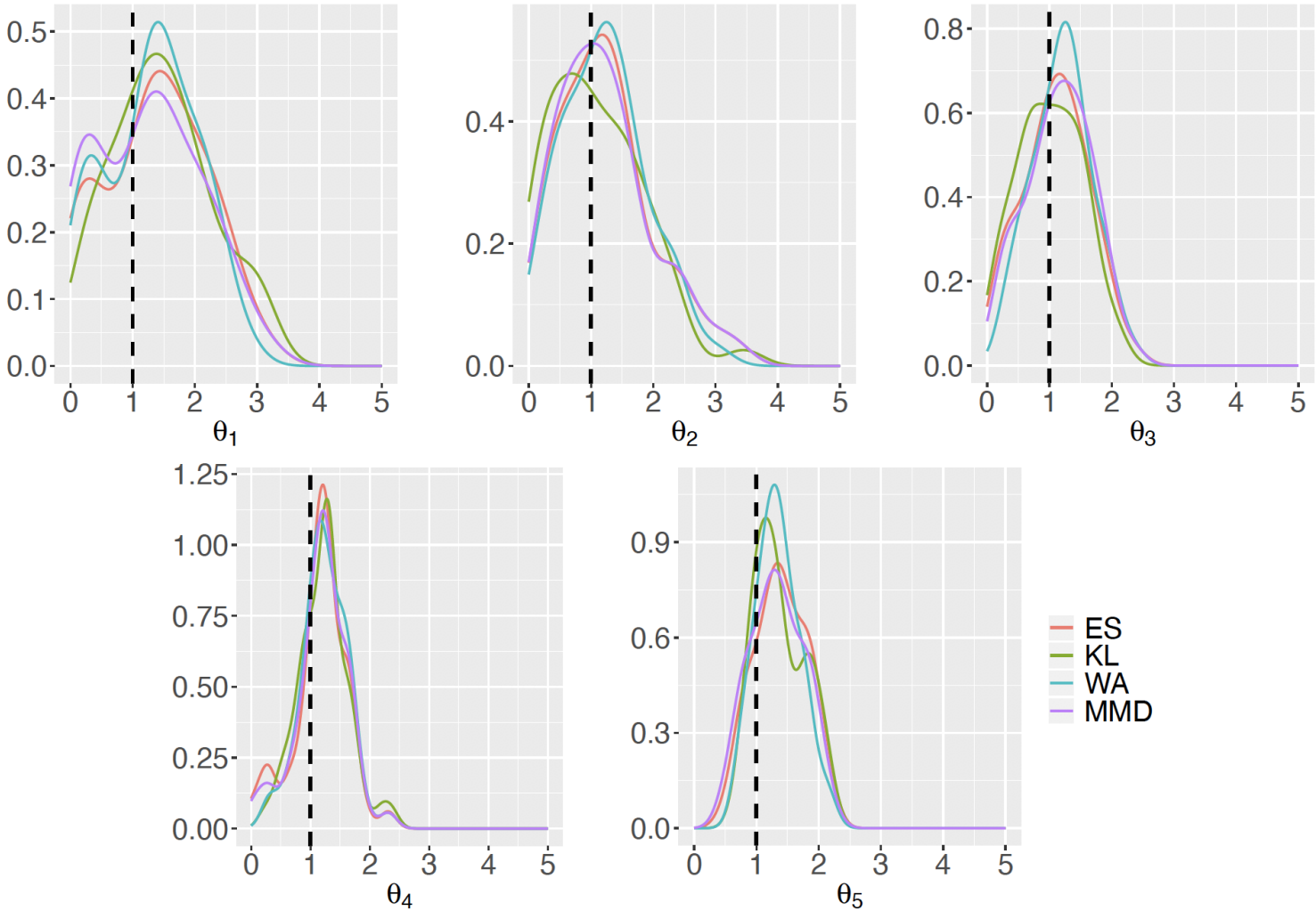
Simulate $\boldsymbol{Y}_{n,k} \in \mathbb{R}^2$ from (2), with $\boldsymbol{\theta}^*$ replaced by $\boldsymbol{\theta}_k$, by simulating $U_{k,j} \sim \text{Gamma}\left(\theta_{k,j}, 1\right)$ and setting

$$Y_{n,k,1} = V_{k,1}/(1 + V_{k,1}), \text{ and } Y_{n,k,2} = V_2/(1 + V_{k,2})$$

where

$$V_{k,1} = \frac{U_{k,1} + U_{k,3}}{U_{k,4} + U_{k,5}}, \text{ and } V_{k,2} = \frac{U_{k,2} + U_{k,4}}{U_{k,3} + U_{k,5}}.$$

Numerical study: bivariate beta distribution

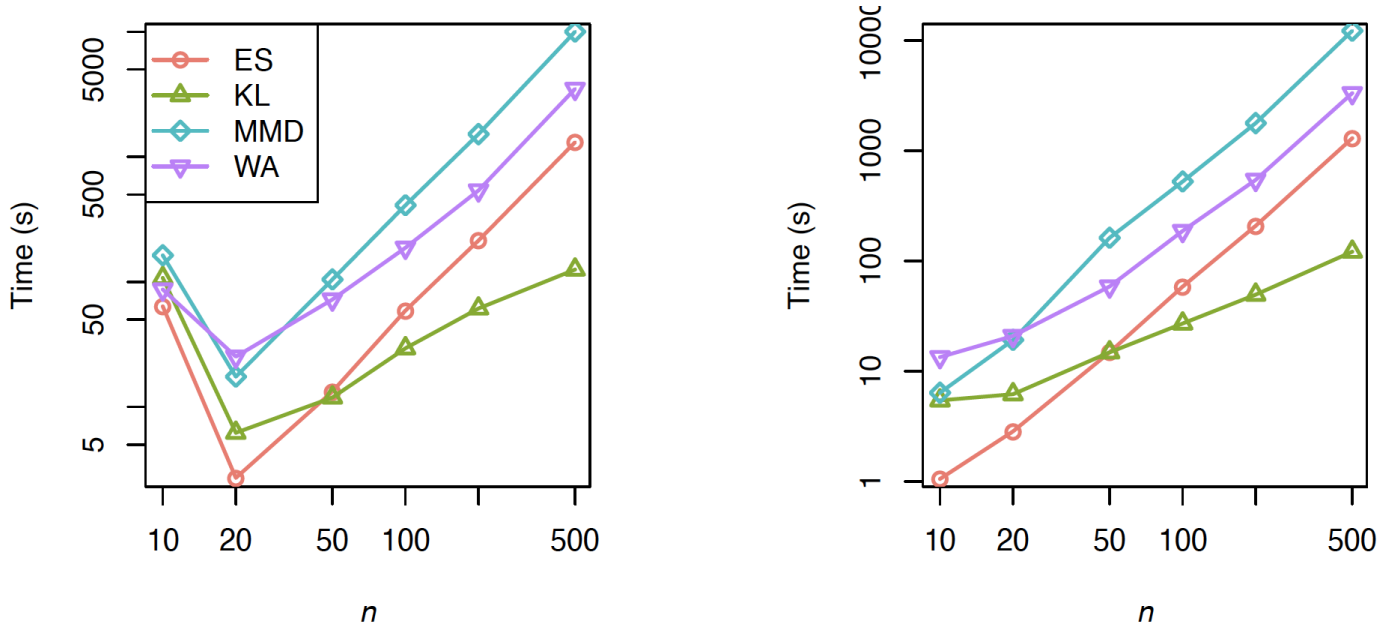


Kernel density estimates of the ABC-obtained posterior samples.

Computational complexities

Discrepancy $\mathscr{D}$	Complexity
All methods (univariate)	$O\left((n+m)\log(n+m)\right)$
KL (multi.)	$O\left((n+m)\log(n+m)\right)$
ES/MMD, approx. WA (multi.)	$O\left((n+m)^2\right)$
WA (multi.)	$O\left((n+m)^{5/2}\log(n+m)\right)$

Timing simulations



Left: results for Gaussian mixture study. **Right:** results for moving average study.

Why not KL?

Jiang et al. (2018) proposed to use the discrepancy

$$\mathscr{D}(\mathbf{x}_n, \mathbf{y}_m) = \frac{q}{n} \sum_{i=1}^{n} \log \left(\frac{\min_{j \in \{1,\dots,n\}} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2}{\min_{j \neq i} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2} \right) + \log \frac{m}{n-1},$$

where $\mathscr{D}(\mathbf{X}_n, \mathbf{Y}_m)$ converges, almost surely, to $\mathrm{KL}(\Pi_X \| \Pi_Y)$.

- Cannot be used for discrete distributions, since $\min_{j \neq i} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = 0$ with non-zero probability.
- Cannot be used for quantized real data, for the same reason.

A linear time estimator for ES/MMD

We can write

$$\mathscr{V}_\delta(\mathbf{x}_n, \mathbf{y}_m) = \sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j_1=1}^{m}\sum_{j_2=1}^{m}\frac{\kappa_\delta(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}; \boldsymbol{y}_{j_1}, \boldsymbol{y}_{j_2})}{m^2 n^2}, \text{ where}$$

$$\begin{aligned}\kappa_\delta(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}; \boldsymbol{y}_{j_1}, \boldsymbol{y}_{j_2}) &= \delta(\boldsymbol{x}_{i_1}, \boldsymbol{y}_{j_1}) + \delta(\boldsymbol{x}_{i_2}, \boldsymbol{y}_{j_2}) \\ &\quad - \delta(\boldsymbol{x}_{i_1}, \boldsymbol{y}_{j_2}) - \delta(\boldsymbol{x}_{i_2}, \boldsymbol{y}_{j_1}).\end{aligned}$$

- When $m = n$, Gretton et al. (2012) observed we can estimate $\mathscr{E}_\delta$, unbiasedly, by

$$\mathscr{U}_\delta(\mathbf{X}_n, \mathbf{Y}_m) = \lfloor n/2 \rfloor^{-1} \sum_{i=1}^{\lfloor n/2 \rfloor} \kappa_\delta(\boldsymbol{X}_{2i-1}, \boldsymbol{X}_{2i}; \boldsymbol{Y}_{2i-1}, \boldsymbol{Y}_{2i}).$$

- We can construct a biased estimator $\mathscr{B}_\delta \geq 0$ that converges to $\mathscr{E}_\delta$, almost surely, by setting

$$\mathscr{B}_\delta(\mathbf{X}_n, \mathbf{Y}_m) = \max\{0, \mathscr{U}_\delta(\mathbf{X}_n, \mathbf{Y}_m)\}.$$

References I

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81:235–269.

Gretton, A., Bogwardt, K. M., Rasch, M. J., Scholkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773.

Jiang, B., Wu, T.-Y., and Wong, W. H. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *Conference on Artificial Intelligence and Statistics (AISTATS)*.

References II

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-ABC: approximate Bayesian computation with kernel embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291.

Szekely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5:1–16.

Szekely, G. J. and Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application*, 4(447-479).

Thank you!

Preprint: **https://arxiv.org/abs/1905.05884**

Email: **h.nguyen5@latrobe.edu.au**

Website: **https://hiendn.github.io**