Distance-based ABC procedures

Hien Nguyen[1]

[1]La Trobe University

(Contact–Email: h.nguyen5@latrobe.edu.au; Website: hiendn.github.io)

ABC in Svalbard (in Melbourne)

LA TROBE UNIVERSITY

Acknowledgements

(a) Florence Forbes.

(b) Julyan Arbel.

(c) Trung Tin Nguyen.

Figure: Highly esteemed collaborators.

Setup

- Let $(\Omega, \mathscr{F}, \Pr)$ be a sufficiently rich probability space and let λ denote the Lebesgue measure.
- Let $\Theta \in \mathbb{T}$ be a random parameter with **prior probability measure** defined by

$$\mathrm{d}\Pi(\theta) = \pi(\theta)\,\mathrm{d}\lambda(\theta).$$

- Let $\boldsymbol{X}_n = (X_i)_{i\in[n]}$ be a random sequence ($[n] = \{1, \ldots, n\}$, $n \in \mathbb{N}$) , where each $X_i \in \mathbb{X}$, and $\boldsymbol{X}_i$, conditioned on $\Theta = \theta$ has probability measure defined by

$$\mathrm{d}F(\boldsymbol{x}_i|\theta) = f(\boldsymbol{x}_i|\theta)\,\mathrm{d}\lambda(\boldsymbol{x}_i).$$

- The joint probability of the pair $W = (\boldsymbol{X}_n, \Theta)$ is thus given by the joint probability measure defined by

$$\mathrm{d}F(\boldsymbol{x}_i, \theta) = \pi(\theta)\, f(\boldsymbol{x}_i|\theta)\,\mathrm{d}\lambda(\boldsymbol{x}_i, \theta).$$

Setup

- Suppose that we observe $\boldsymbol{X}_n = \boldsymbol{x}_n$ from a pair W, but not Θ, which we wish to estimate.
- Using the previous characterizations, we can draw inference regarding Θ, conditionally on $\boldsymbol{X}_n = \boldsymbol{x}_n$ by constructing the **posterior PDF**

$$\pi(\theta|\boldsymbol{x}_n) = \frac{\pi(\theta) f(\boldsymbol{x}_i|\theta)}{c(\boldsymbol{x}_i)},$$

  where $c(\boldsymbol{x}_i) = \int_{\mathbb{T}} \pi(\theta) f(\boldsymbol{x}_i|\theta) \mathrm{d}\lambda(\theta)$.
- Often $f(\boldsymbol{x}_i|\theta)$ is not available in a tractable form, but it is still possible to simulate from $f(\boldsymbol{x}_i|\theta)$.

Distance-based ABC

- For two sequences $\boldsymbol{x}_n, \boldsymbol{y}_n \in \mathbb{X}^n$, define the **discrepancy** $D(\boldsymbol{x}_n, \boldsymbol{y}_n) \geq 0$.
- For $d \geq 0$ and $\varepsilon > 0$, define the **weight function** $w(d, \varepsilon)$, where $w(\cdot, \varepsilon)$ is decreasing in d.
- From a simulation $\boldsymbol{W}_N = (W_j)_{j \in [N]}$, where $W_j = (\boldsymbol{Y}_{n,j}, \Theta_j)$ and $N \in \mathbb{N}$, from $f(\boldsymbol{y}_i, \theta)$, we can estimate the characteristics of the so called **ABC pseudo-posterior PDF**

$$\pi_\varepsilon(\theta|\boldsymbol{x}_n) = \frac{\pi(\theta) L_\varepsilon(\boldsymbol{x}_n|\theta)}{c_\varepsilon(\boldsymbol{x}_n)},$$

where

$$L_\varepsilon(\boldsymbol{x}_n|\theta) = \int_{\mathbb{X}^n} w(D(\boldsymbol{x}_n, \boldsymbol{y}_n), \varepsilon) f(\boldsymbol{y}_n|\theta) \mathrm{d}\lambda(\boldsymbol{y}_n)$$

and $c_\varepsilon(\boldsymbol{x}_n) = \int_{\mathbb{T}} \pi(\theta) L_\varepsilon(\boldsymbol{x}_n|\theta) \mathrm{d}\lambda(\theta)$.

Discrepancies

- $D(\boldsymbol{x}_n, \boldsymbol{y}_n) = \|s(\boldsymbol{x}_n) - s(\boldsymbol{y}_n)\|_p$, $p \geq 1$, where $s(\cdot)$ is a **summary statistic**.
- Jiang et al. (2018) considered the **sample Kullback–Leibler divergence** of Perez-Cruz (2008), for $\mathbb{X} \subseteq \mathbb{R}^d$:

$$D(\boldsymbol{x}_n, \boldsymbol{y}_n) = \frac{d}{n} \sum_{i=1}^{n} \log \frac{\min_j \|x_i - y_j\|_2}{\min_{j \neq i} \|x_i - y_j\|_2} + \log \frac{n}{n-1}.$$

- Bernton et al. (2019) considered $D(\boldsymbol{x}_n, \boldsymbol{y}_n) = \mathfrak{W}_p(\boldsymbol{x}_n, \boldsymbol{y}_n)$, where $\mathfrak{W}_p$ is the p-**Wasserstein distance** between multisets of n vectors.
- Nguyen et al. (2020) considered the **energy statistic** of Szekely and Rizzo (2004), for $\mathbb{X} \subseteq \mathbb{R}^d$:

$$D(\boldsymbol{x}_n, \boldsymbol{y}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ 2\|x_i - y_j\|_2 - \|x_i - x_j\|_2 - \|y_i - y_j\|_2 \right\}.$$

Weight functions

- Typically we choose the weight function w to be

$$w(d, \varepsilon) = \mathbf{1}\{d \leq \varepsilon\},$$

where $\mathbf{1}\{\cdot\}$ is the indicator function.

  - Here, the choice corresponds to the traditional **rejection-sampling** ABC algorithms.

- A more exotic choice is considered in Park et al. (2016) and Miller and Dunson (2019), where

$$w(d, \varepsilon) = \exp\left(-\varepsilon^{-1} d^p\right),\ p \geq 1.$$

  - This corresponds to **importance sampling**-based ABC procedures, as considered in Karabatsos and Leisen (2018).

Behavior of the pseudo-posterior when $\varepsilon \to 0$

A1 Use $w(d, \varepsilon) = \mathbf{1}\{d \leq \varepsilon\}$.

A2 The PDF $f(\boldsymbol{x}_n|\theta)$ is continuous and uniformly bounded, for all $\theta \in \mathbb{T}$ and $\boldsymbol{x}_n \in \mathbb{X}^n$.

A3 The discrepancy is continuous in the sense that $D(\boldsymbol{x}_n, \boldsymbol{y}_n) \to D(\boldsymbol{x}_n, \bar{\boldsymbol{y}}_n)$, when $\boldsymbol{y}_n \to \bar{\boldsymbol{y}}_n$.

A4 $\mathrm{d}F(\boldsymbol{x}_n|\theta)$ defines an exchangeable measure and $D(\boldsymbol{x}_n, \boldsymbol{y}_n) = 0$, if and only if $\{x_i\}_{i=1}^n = \{y_i\}_{i=1}^n$.

The following result is proved by Rubio and Johansen (2013) and Bernton et al. (2019).

Proposition 1

Under A1–A4, for fixed $\boldsymbol{x}_n \in \mathbb{X}_n$,

$\int_{\mathbb{T}} |\pi_\varepsilon(\theta|\boldsymbol{x}_n) - \pi(\theta|\boldsymbol{x}_n)| \, d\lambda(\theta) \to 0$, *as* $\varepsilon \to 0$.

Behavior of the pseudo-posterior when $n \to \infty$

B1 $\boldsymbol{X}_n$ and $\boldsymbol{Y}_n$ have probability measures defined by $\mathrm{d}F(\boldsymbol{x}_n|\theta_0)$ and $\mathrm{d}F(\boldsymbol{y}_n|\theta)$, respectively.

B2 $D(\boldsymbol{X}_n, \boldsymbol{Y}_n) \to D(\theta_0, \theta)$, almost surely, as $n \to \infty$, where $D(\theta_0, \theta)$ is a function of $\theta_0, \theta \in \mathbb{T}$.

B3 $w(d, \varepsilon)$ is piecewise continuous, and $w(d, \varepsilon) \leq a < \infty$.

The following is proved by Jiang et al. (2018), Miller and Dunson (2019) and Nguyen et al. (2020).

Theorem 1

Under B1–B3, for any $\varepsilon > 0$, if $w(\cdot, \varepsilon)$ is continuous at $D(\theta_0, \theta)$, then

$$\pi_\varepsilon(\theta|\boldsymbol{X}_n) \xrightarrow[n\to\infty]{a.s.} \frac{\pi(\theta)\, w(D(\theta_0, \theta), \varepsilon)}{\int_{\mathbb{T}} \pi(\theta)\, w(D(\theta_0, \theta), \varepsilon)\, d\lambda(\theta)} = \pi_\varepsilon(\theta; \theta_0).$$

Behavior of the pseudo-posterior when $n \to \infty$

- Let $Q(u;\omega)$ and $Q_n(u;\omega)$ $(n \in \mathbb{N})$ be probability measures with respect to $u \in (\mathbb{U}, \mathscr{U})$, for each $\omega \in \Omega$.
- By Berti et al. (2006) and Grubel and Kabluchko (2016), say that the sequence Q_n **almost surely weakly** (a.s.w) converges to Q, if on a set $A \in \mathscr{F}$ with $\Pr(A) = 1$, for all $\omega \in A$, $Q_n(\cdot;\omega)$ weakly converges on $\mathbb{U}$ to $Q(\cdot;\omega)$, as $n \to \infty$.
- Consider the pseudo-posterior and its limiting measure, defined by

$$d\Pi_\varepsilon(\theta|\boldsymbol{X}_n) = \pi_\varepsilon(\theta|\boldsymbol{X}_n)\, d\lambda(\theta),\ d\Pi_\varepsilon(\theta;\theta_0) = \pi_\varepsilon(\theta;\theta_0)\, d\lambda(\theta).$$

Corollary 1

Under the assumptions of Theorem 1, if $\mathbb{T} \subseteq \mathbb{R}^q$, $q \in \mathbb{N}$, then $\Pi_\varepsilon(\cdot|\boldsymbol{X}_n)$ a.s.w. converges on $\mathbb{T}$ to $\Pi_\varepsilon(\cdot;\theta_0)$, as $n \to \infty$.

Behavior of the pseudo-posterior when $n \to \infty$

C1 Use $w(d, \varepsilon) = \mathbf{1}\{d \leq \varepsilon\}$.

C2 $D(\theta_0, \theta)$ is a metric on the set $\mathbb{T} \subseteq \mathbb{R}^q$, $q \in \mathbb{N}$.

C3 The identity function $\text{id} : (\mathbb{T}, D) \to \left(\mathbb{T}, \|\cdot - \cdot\|_p\right)$ is uniformly continuous, for some $p \geq 1$ (satisfied if $\|\theta_0 - \theta\|_p \leq bD(\theta_0, \theta)$, for some finite constant $b > 0$).

Corollary 2

Assume C1–C3 and the assumptions of Corollary 1. For any $\iota > 0$, there exists an $\varepsilon > 0$ (independent of θ_0), and a set $\mathbb{S}_\iota(\theta_0) \subset \mathbb{T}$ containing θ_0, with $\lambda(\mathbb{S}_\iota(\theta_0)) = \iota$, such that

$$\liminf_{n \to \infty} \Pi_\varepsilon(\mathbb{S}_\iota(\theta_0) | \boldsymbol{X}_n) = 1, \text{ a.s.}$$

Illustration

- $\boldsymbol{X}_n$ and $\boldsymbol{Y}_n$ are IID sequences, such that $X_i \sim \mathrm{N}\left(\theta_0, \sigma^2\right)$ and $Y_i \sim \mathrm{N}\left(\theta, \sigma^2\right)$, with $\sigma^2 > 0$ fixed.
- $\theta \sim \mathrm{N}\left(0, \tau^2\right)$, with $\tau^2 > 0$ fixed.
- Use $w\left(d, \varepsilon\right) = \mathbf{1}\left\{d \leq \varepsilon\right\}$.
- Use $D\left(\boldsymbol{X}_n, \boldsymbol{Y}_n\right) = \left| n^{-1} \sum_{i=1}^{n} X_i - n^{-1} \sum_{i=1}^{n} Y_i \right| \xrightarrow[n \to \infty]{\text{a.s.}} \left|\theta_0 - \theta\right|$.

Illustration

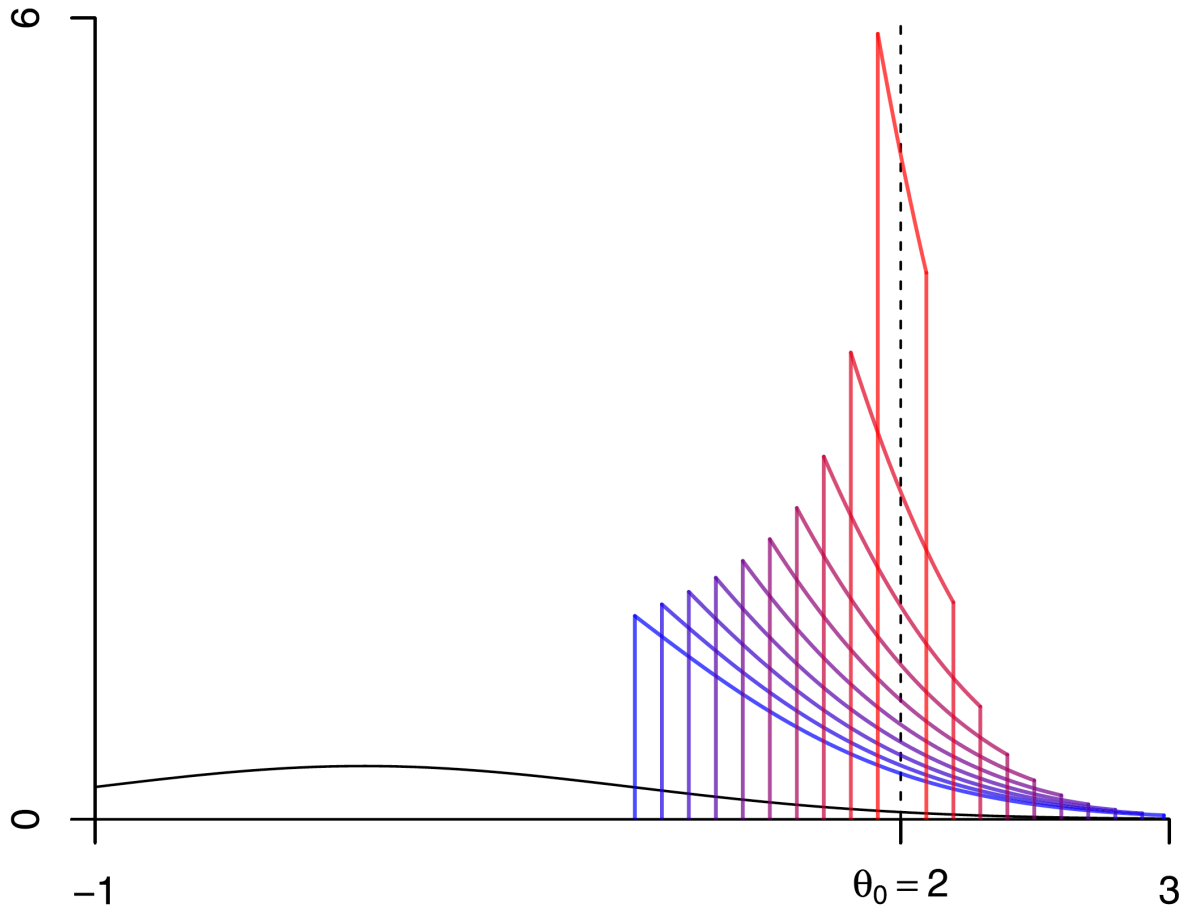

Figure: Limiting pseudo-posteriors $\pi_\varepsilon(\theta;\theta_0)$ with $\varepsilon \in [0.1, 1]$, when $\theta_0 = 2$. The black curve represents the prior density, and the red–blue curves represent the pseudo-posteriors for increasing ε.

The energy distance

- Recall that for $\mathbb{X} \subseteq \mathbb{R}^d$, we define the **energy statistic** between two samples $\boldsymbol{x}_n$ and $\boldsymbol{y}_n$ as:

$$D(\boldsymbol{x}_n, \boldsymbol{y}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ 2 \left\| x_i - y_j \right\|_2 - \left\| x_i - x_j \right\|_2 - \left\| y_i - y_j \right\|_2 \right\}.$$

- From Szekely and Rizzo (2017), we have the fact $\sqrt{D(\boldsymbol{x}_n, \boldsymbol{y}_n)}$ is a metric on multisets of n vectors.
- From Nguyen et al. (2020), for IID $\boldsymbol{X}_n$ and $\boldsymbol{Y}_n$ with parameters θ_0 and θ, and $\mathrm{E}\left[\left\| X_1 \right\|_2^2\right] + \mathrm{E}\left[\left\| Y_1 \right\|_2^2\right] < \infty$, we have $D(\boldsymbol{X}_n, \boldsymbol{Y}_n)$ a.s. converges, as $n \to \infty$, to

$$D(\theta_0, \theta) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\pi^{\frac{d+1}{2}}} \int_{\mathbb{R}^d} \frac{\left| \mathrm{E} \exp\left(\mathrm{i} t^\top X_1\right) - \mathrm{E} \exp\left(\mathrm{i} t^\top Y_1\right) \right|^2}{\left\| t \right\|_2^{d+1}} \mathrm{d}\lambda(t).$$

The energy distance

Write $\boldsymbol{Z}_n = (Z_i)_{i\in[n]}$, where $Z_i = (X_i, Y_i)$.

D1 $\boldsymbol{Z}_n$ is a contiguous subsequence of $\boldsymbol{Z}_{\mathbb{Z}} = (Z_i)_{i\in\mathbb{Z}}$.

D2 The sequence $\boldsymbol{Z}_{\mathbb{Z}}$ is stationary and ergodic, where $\boldsymbol{X}_{\mathbb{Z}}$ and $\boldsymbol{Y}_{\mathbb{Z}}$ are independent.

D3 X_1 and Y_1 have marginal measures defined by $\mathrm{d}F_1(x|\theta_0)$ and $\mathrm{d}F_1(y|\theta)$, and $\mathrm{E}\,\|X_1\|_2 + \mathrm{E}\,\|Y_1\|_2 < \infty$.

We have the following result, using the ergodic U-statistics theorem of Aaronson et al. (1996).

Proposition 2

Under assumptions D1–D3, $D(\boldsymbol{X}_n, \boldsymbol{Y}_n)$ a.s. converges, as $n \to \infty$, to $D(\theta_0, \theta)$.

Example application

- Let $\boldsymbol{X}_n$ arise from an ARMA$(1,1)$ model with $\theta_0 = (\mathrm{AR}_0, \mathrm{MA}_0)$ with standard normal noise, $n = 1000$.
- Let $\boldsymbol{Y}_n$ arise from an ARMA$(1,1)$ model with $\Theta \sim \text{Uniform}\left([-1,1]^2\right)$ with standard normal noise.
- Construct the lagged vector sequences $\tilde{\boldsymbol{X}}_n = ((X_i, X_{i+1}))_{i \in [n-1]}$ and $\tilde{\boldsymbol{Y}}_n = ((Y_i, Y_{i+1}))_{i \in [n-1]}$.
- Generate $\boldsymbol{W}_N = ((\boldsymbol{Y}_{n,j}, \Theta_j))_{j \in [N]}$ consisting of $N = 10000$ realizations.
- Compute the energy statistics $\boldsymbol{D}_N = \left(D\left(\tilde{\boldsymbol{X}}_n, \tilde{\boldsymbol{Y}}_{n,j}\right)\right)_{j \in [N]}$.
- Determine the sets of $\boldsymbol{D}_N$ that satisfy $D\left(\tilde{\boldsymbol{X}}_n, \tilde{\boldsymbol{Y}}_{n,j}\right) < \varepsilon$, for various levels of ε, determined by the quantiles of $\boldsymbol{D}_N$.

Example application

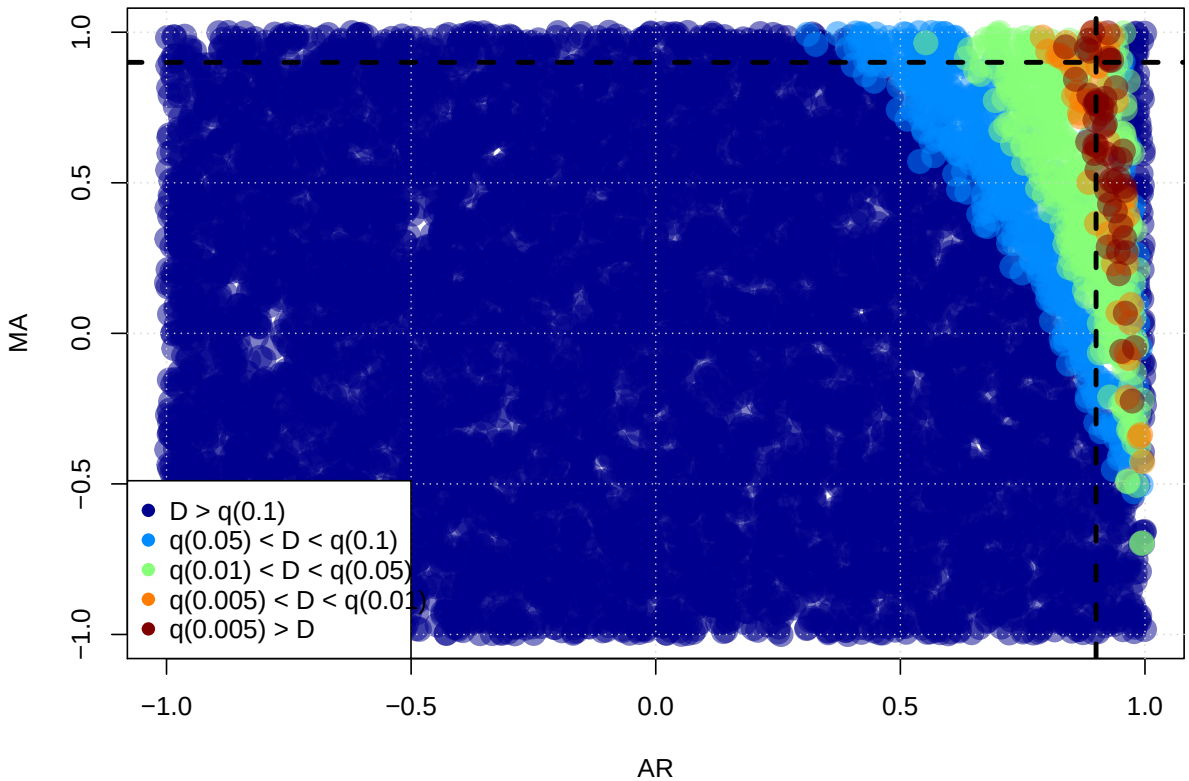

Figure: $\theta_0 = (\mathrm{AR}_0, \mathrm{MA}_0) = (0.9, 0.9)$.

Example application

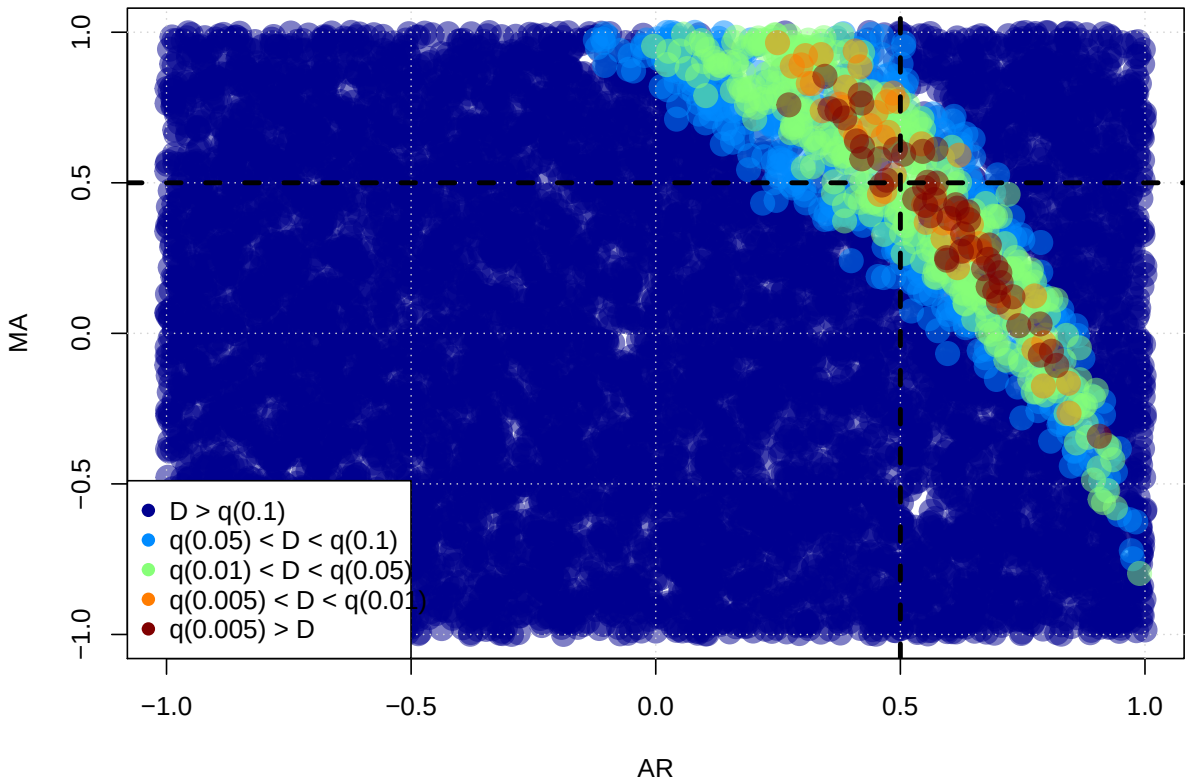

Figure: $\theta_0 = (\mathrm{AR}_0, \mathrm{MA}_0) = (0.5, 0.5)$.

Example application

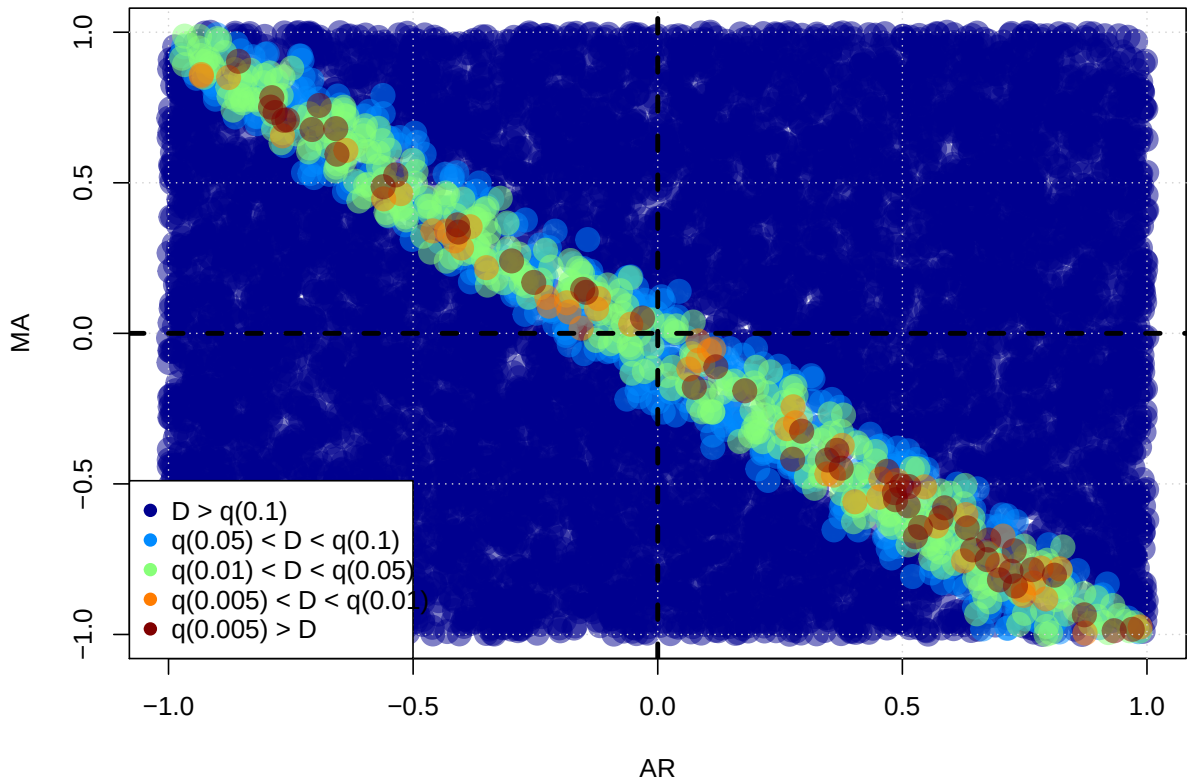

Figure: $\theta_0 = (\mathrm{AR}_0, \mathrm{MA}_0) = (0, 0)$.

A better example

- Let $\boldsymbol{X}_n$ arise from an ARMA$(0,2)$ model with $\theta_0 = (\text{MA1}_0, \text{MA2}_0)$ with standard normal noise, $n = 1000$.
- Let $\boldsymbol{Y}_n$ arise from an ARMA$(0,2)$ model with $\Theta \sim \text{Uniform}\left([-1,1]^2\right)$ with standard normal noise.
- Construct the lagged vector sequences $\tilde{\boldsymbol{X}}_n = ((X_i, X_{i+1}))_{i \in [n-1]}$ and $\tilde{\boldsymbol{Y}}_n = ((Y_i, Y_{i+1}))_{i \in [n-1]}$.
- Generate $\boldsymbol{W}_N = ((\boldsymbol{Y}_{n,j}, \Theta_j))_{j \in [N]}$ consisting of $N = 10000$ realizations.
- Compute the energy statistics $\boldsymbol{D}_N = \left(D\left(\tilde{\boldsymbol{X}}_n, \tilde{\boldsymbol{Y}}_{n,j}\right)\right)_{j \in [N]}$.
- Determine the sets of $\boldsymbol{D}_N$ that satisfy $D\left(\tilde{\boldsymbol{X}}_n, \tilde{\boldsymbol{Y}}_{n,j}\right) < \varepsilon$, for various levels of ε, determined by the quantiles of $\boldsymbol{D}_N$.

A better example

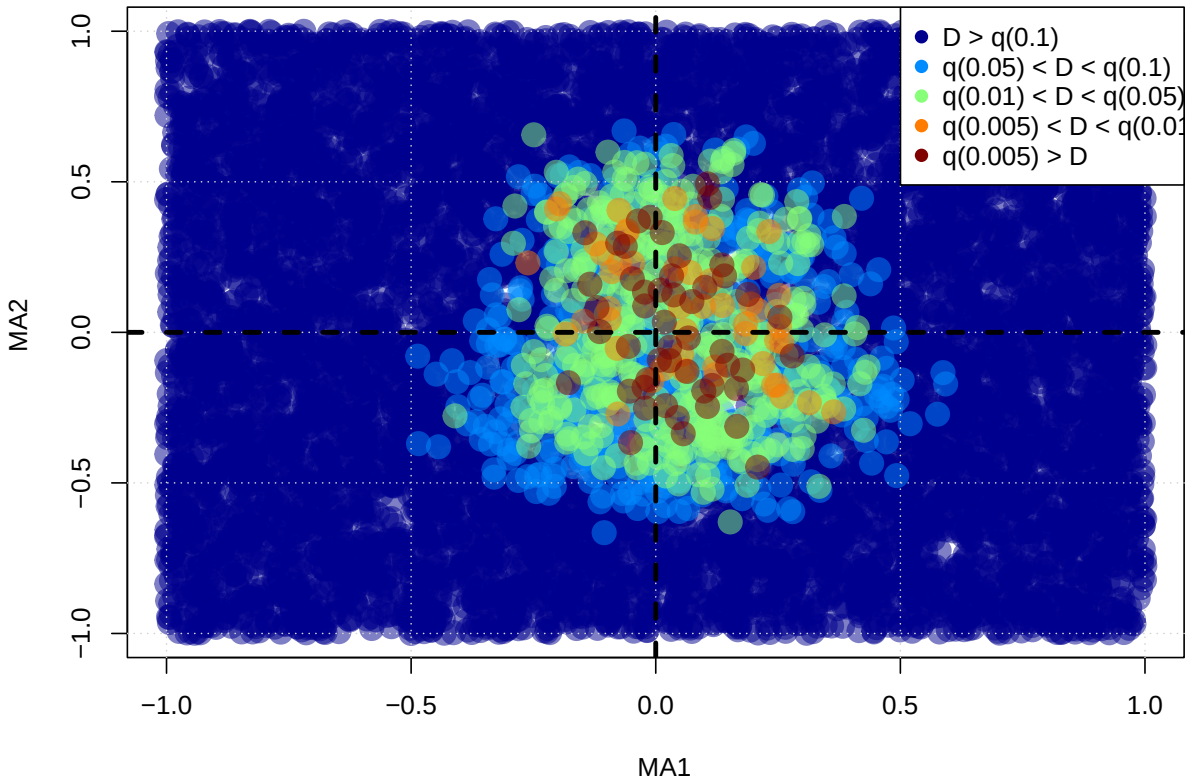

Figure: $\theta_0 = (\mathrm{MA1}_0, \mathrm{MA2}_0) = (0, 0)$.

A better example

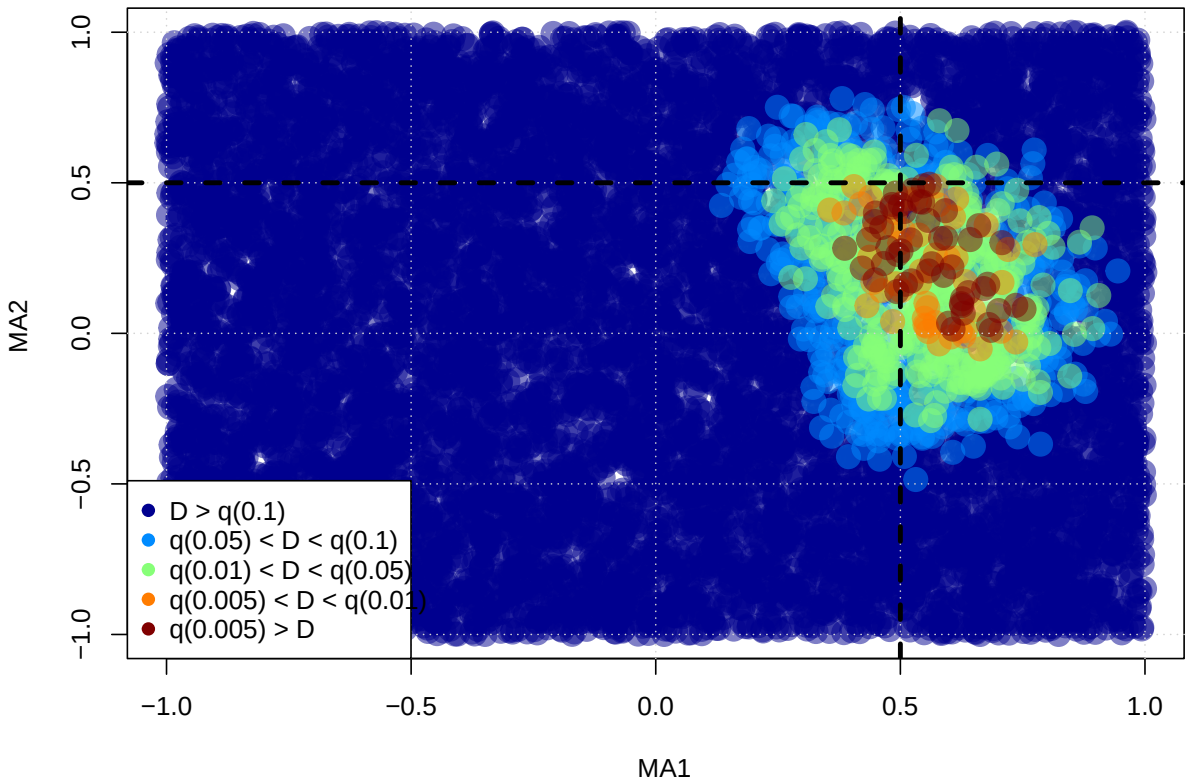

Figure: $\theta_0 = (\mathrm{MA1}_0, \mathrm{MA2}_0) = (0.5, 0.5)$.

A better example

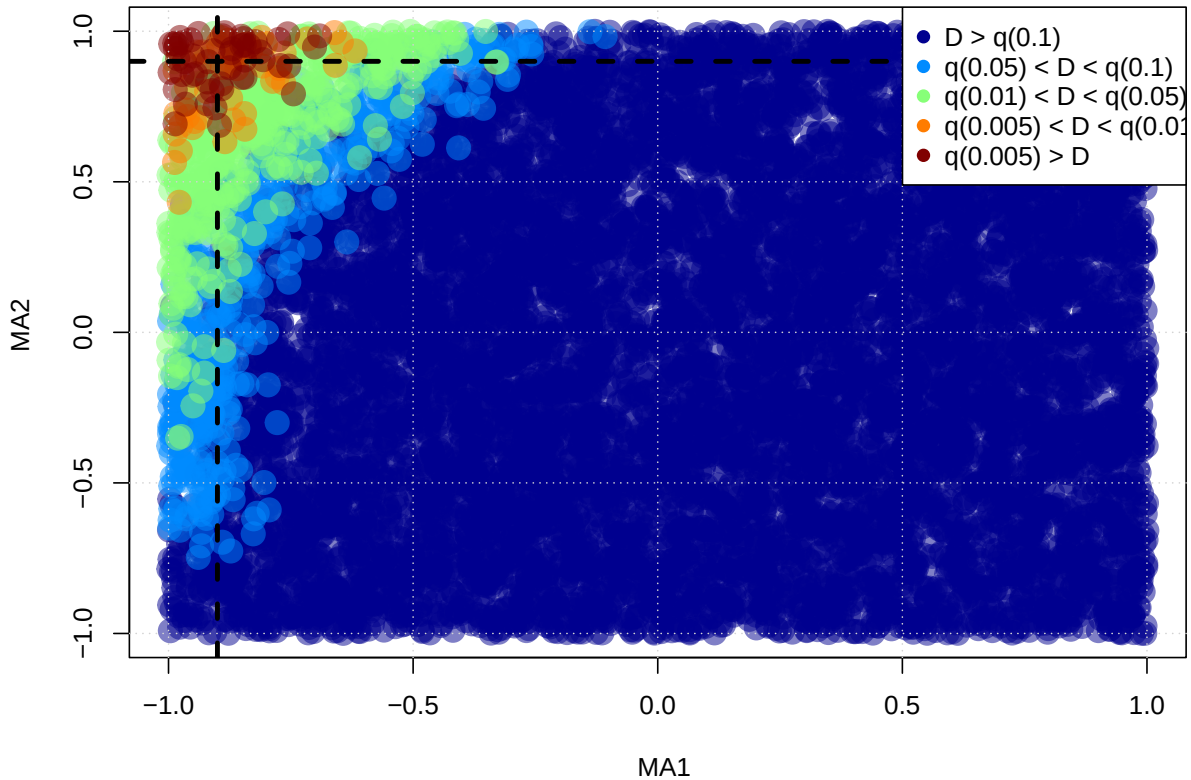

Figure: $\theta_0 = (\text{MA1}_0, \text{MA2}_0) = (-0.9, 0.9)$.

References I

Aaronson, J., Burton, R., Dehling, H., Gilat, D., Hill, T., and Weiss, B. (1996). Strong laws for L- and U-statistics. *Transactions of the American Mathematical Society*, 348:2845–2866. 15

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 B*, 8:235–269. 6, 8

Berti, P., Pratelli, L., and Rigo, P. (2006). Almost sure weak convergence of random probability measures. *Stohastics and Stochastics Reports*, 78:91–97. 10

Grubel, R. and Kabluchko, Z. (2016). A functional central limit theorem for branching random walks, almost sure weak convergence and applications to random trees. *Annals of Applied Probability*, 26:3659–3698. 10

References II

Jiang, B., Wu, T.-Y., and Wong, W. H. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *AISTATS*. 6, 9

Karabatsos, G. and Leisen, F. (2018). An approximate likelihood perspective on ABC methods. *Statistics Surveys*, 12:66–104. 7

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114:1113–1125. 7, 9

Nguyen, H. D., Arbel, J., Lu, H., and Frobes, F. (2020). Approximate Bayesian computation via the energy statistic. *IEEE Access*, 8:131683–131698. 6, 9, 14

References III

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-ABC: approximate Bayesian computation with kernel embeddings. In *AISTATS*. 7

Perez-Cruz, F. (2008). Kullback-Leibler divergence estimation of continuous distributions. In *ISIT 2008*. 6

Rubio, F. J. and Johansen, A. M. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, 7:1632–1654. 8

Szekely, G. J. and Rizzo, M. L. (2004). Testing for equal distributions in high dimension. *InterStat*, 5:1249–1272. 6

Szekely, G. J. and Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistis and Its Application*, 4:447–479. 14

Thank you!

Most recent paper:

https://doi.org/10.1109/ACCESS.2020.3009878

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

Website: **hiendn.github.io**