Kernel Recursive ABC: Point Estimation with Intractable Likelihood
Supplementary Materials

A Proofs for theoretical results

We here provide proofs for the theoretical results in Section 3 of the main text. For ease of understanding, we repeat the assumptions and the statements w.r.t. those results. The notation follows that of the main text.

Assumption 1. *(i) ℓ has a unique global maximum at $\theta_\infty \in \Theta$, and $\pi(\theta_\infty) > 0$; (ii) π is continuous at θ_∞, ℓ has continuous second derivatives in the neighborhood of θ_∞, and the Hessian of ℓ at θ_∞ is strictly negative-definite.*

Proposition 1. *Let $\Theta \subset \mathbb{R}^d$ be a Borel measurable set, $k : \Theta \times \Theta \to \mathbb{R}$ be a continuous, bounded kernel, and $\mathcal{H}$ be its RKHS. If Assumption 1 holds, then we have*

$$\lim_{N\to\infty} \|\mu_{P_N} - k(\cdot, \theta_\infty)\|_{\mathcal{H}} = 0.$$

Proof. Because Assumption 1 is equivalent to Assumptions A1, A2 and A3 in Lele *et al.* (2010), we can use the Corollary to Lemma A.2 on p.1624 of Lele *et al.* (2010); this guarantees the weak convergence of P_N to δ_{θ_∞}, the Dirac distribution at θ_∞. Therefore,

$$\begin{aligned}
\lim_{N\to\infty} \|\mu_{P_N} - k(\cdot, \theta_\infty)\|^2_{\mathcal{H}} &= \lim_{N\to\infty} \langle \mu_{P_N}, \mu_{P_N} \rangle_{\mathcal{H}} - 2 \lim_{N\to\infty} \langle \mu_{P_N}, k(\cdot, \theta_\infty) \rangle_{\mathcal{H}} \\
&\quad + \langle k(\cdot, \theta_\infty), k(\cdot, \theta_\infty) \rangle_{\mathcal{H}} \\
&= \lim_{N\to\infty} \int\int k(\theta, \theta') dP_N(\theta) dP_N(\theta') \\
&\quad -2 \lim_{N\to\infty} \int k(\theta, \theta_\infty) dP_N(\theta) + k(\theta_\infty, \theta_\infty) \\
&= k(\theta_\infty, \theta_\infty) - 2k(\theta_\infty, \theta_\infty) + k(\theta_\infty, \theta_\infty) \qquad (1) \\
&= 0,
\end{aligned}$$

where (1) follows from the weak convergence of P_N to δ_{θ_∞} and k is continuous and bounded. Here we have used Theorem 2.8 (ii) in Billingsley (1999) for the first term in (1). □

Assumption 2. *(i) There exists a constant $C > 0$ such that $k(\theta, \theta) = C$ for all $\theta \in \Theta$. (ii) It holds that $k(\theta, \theta') < C$ for all $\theta, \theta' \in \Theta$ with $\theta \neq \theta'$.*

Proposition 2. *Let $\Theta \subset \mathbb{R}^d$ be a compact set and $k : \Theta \times \Theta \to \mathbb{R}$ be a continuous, bounded kernel. Let $\theta_N := \operatorname{argmin}_{\tilde{\theta}\in\Theta} \left\|\mu_{P_N} - k(\cdot, \tilde{\theta})\right\|_{\mathcal{H}}$. If Assumptions 1 and 2 hold, then we have $\theta_N \to \theta_\infty$ as $N \to \infty$.*

Proof. By the reproducing property and Assumption 2, we have

$$\begin{aligned} \|\mu_{P_N} - k(\cdot, \tilde{\theta})\|_{\mathcal{H}}^2 &= \|\mu_{P_N}\|_{\mathcal{H}}^2 - 2\mu_{P_N}(\tilde{\theta}) + k(\tilde{\theta}, \tilde{\theta}) \\ &= \|\mu_{P_N}\|_{\mathcal{H}}^2 - 2\int k(\tilde{\theta}, \theta) dP_N(\theta) + C. \end{aligned}$$

Since $\int k(\tilde{\theta}, \theta) dP_N(\theta)$ is a continuous function of $\tilde{\theta}$ (which follows from the continuity of k and the dominated convergence theorem), it follows that $\|\mu_{P_N} - k(\cdot, \tilde{\theta})\|_{\mathcal{H}}^2$ is a continuous function of $\tilde{\theta}$, and so is $\|\mu_{P_N} - k(\cdot, \tilde{\theta})\|_{\mathcal{H}}$. Thus, since Θ is compact, $\theta_N = \operatorname{argmin}_{\tilde{\theta}\in\Theta} \left\|\mu_{P_N} - k(\cdot, \tilde{\theta})\right\|_{\mathcal{H}}$ exists. Using the above identity, we then have

$$\begin{aligned} \theta_N &= \operatorname*{argmin}_{\tilde{\theta}\in\Theta} \left\|\mu_{P_N} - k(\cdot, \tilde{\theta})\right\|_{\mathcal{H}}^2 \\ &= \operatorname*{argmin}_{\tilde{\theta}\in\Theta} \|\mu_{P_N}\|_{\mathcal{H}}^2 - 2\mu_{P_N}(\tilde{\theta}) + C \\ &= \operatorname*{argmax}_{\tilde{\theta}\in\Theta} \mu_{P_N}(\tilde{\theta}). \end{aligned}$$

By the reproducing property, the Cauchy-Schwartz inequality, and Assumption 2, we have for all $\theta \in \Theta$

$$\begin{aligned} |\mu_{P_N}(\theta) - k(\theta, \theta_\infty)| &= |\langle k(\cdot, \theta), \mu_{P_N} - k(\cdot, \theta_\infty)\rangle| \\ &\leq \sqrt{k(\theta, \theta)}\, \|\mu_{P_N} - k(\cdot, \theta_\infty)\|_{\mathcal{H}} \\ &= \sqrt{C}\, \|\mu_{P_N} - k(\cdot, \theta_\infty)\|_{\mathcal{H}} \end{aligned} \tag{2}$$

Let ε be an arbitrary positive number and $U_\varepsilon(\theta_\infty)$ be an open ε-neighborhood of θ_∞. From Assumption 2 (ii) and the continuity of k, there is $\delta > 0$ such that

$$\max_{\theta\in\Theta\setminus U_\varepsilon(\theta_\infty)} k(\theta, \theta_\infty) \leq C - \delta. \tag{3}$$

It follows from Eq.(2) and Proposition 1 that there is $N_0 \in \mathbb{N}$ such that

$$\max_{\theta\in\Theta} |\mu_{P_N}(\theta) - k(\theta, \theta_\infty)| \leq \delta/3 \tag{4}$$

holds for all $N \geq N_0$. This implies, in particular, that for all $N \geq N_0$

$$\mu_{P_N}(\theta_\infty) \geq k(\theta_\infty, \theta_\infty) - \delta/3 = C - \delta/3. \tag{5}$$

On the other hand, using Eqs.(3) and (4), we have

$$\max_{\theta\in\Theta\setminus U_\varepsilon(\theta_\infty)} \mu_{P_N}(\theta) \leq C - \frac{2}{3}\delta \tag{6}$$

for all $N \geq N_0$.

Eqs.(5) and (6) show that the maximum of μ_{P_N} is attained in $U_\varepsilon(\theta_\infty)$, that is, $\theta_N \in U_\varepsilon(\theta_\infty)$, for all $N \geq N_0$, which completes the proof. □

Remark 1. *For simplicity, we assume in Proposition 2 that θ is compact, but this condition can be relaxed. For example, we may instead assume the following weaker condition: For any open neighborhood U of θ_∞, there is a positive constant δ such that* $\sup_{\theta \in \Theta \setminus U} k(\theta, \theta_\infty) \leq k(\theta_\infty, \theta_\infty) - \delta$.

B Demonstration of the auto-correction mechanism for a misspecified prior

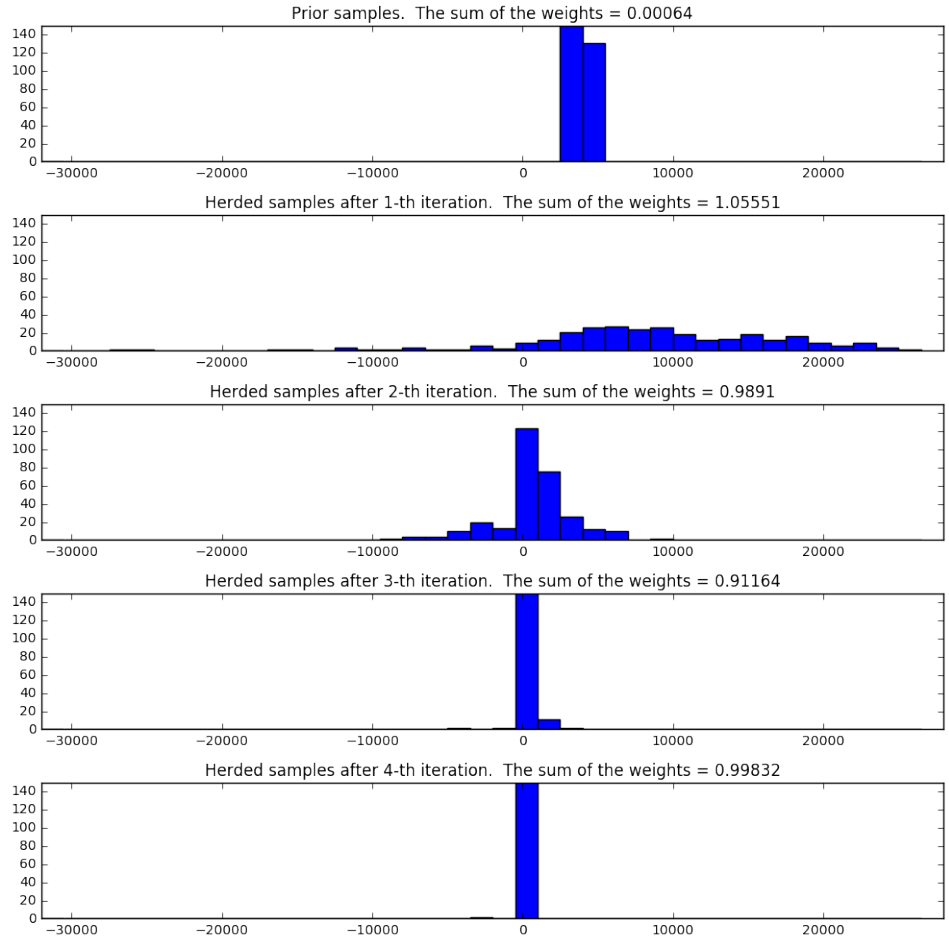

Figure 1: Each figure shows a histogram of simulated parameters for the mean of the Gaussian distribution in each iteration, as produced with the proposed method. "The sum of the weights" on the top of each figure is the sum of the weights given by kernel ABC at each iteration, as defined by Eq. (3) of the main text.

We demonstrate here how the auto-correction mechanism of the proposed method works; for an explanation of this mechanism, see Section 3 of the main text. We performed an experiment similar to the one in Section 4.2 of the main text, but under a simpler setting. The task was to estimate the mean

0 of a univariate Gaussian distribution, Normal$(0, 40)$, provided 100 i.i.d. observations from it. The variance 40 was assumed to be known. For the prior distribution over the mean, we used the uniform distribution on $[2000, 3000]$, which is severely misspecified. For the proposed method, we recomputed the bandwidth of a Gaussian kernel for each iteration, by using the median heuristic with simulated data. In each iteration, 300 pseudo-observations were generated for the proposed method.

Figure 1 shows the results for the first 4 iterations. The top figure is a histogram of the parameters generated from the prior distribution, which, because of the misspecification of the prior, do not cover the true mean 0. The resulting sum of the weights is 0.00064, implying that the simulated pseudo-observations are far apart from the observed data. (As explained in the caption of Figure 1, "The sum of the weights" on the top of each figure is the sum of the weights $w_1, \ldots, w_n$ given by kernel ABC at each iteration, as defined by Eq. (3) of the main text.) The second figure is a histogram of the parameters generated by kernel herding in the first iteration. These parameters were generated so as to explore the parameter space, in response to the auto-correction mechanism explained in Section 3 of the main text. Since the simulated parameters were now scattered around the true mean 0, kernel ABC began to perform well from the next iteration. After only 4 iterations, the simulated parameters concentrated around the true mean.

C Supplementary to the population dynamics experiment in Section 4.3

We offer here supplementary materials for the experiment on the blowfly population dynamics in Section 4.3 of the main text.

C.1 Errors for individual parameters

Table 1: Results for blowfly population dynamics in Sec. 4.3

Algorithm	P	N_0	σ_d	σ_p	τ	δ	data	cputime
KR-ABC	0.28(0.13)	0.03(0.05)	0.93(0.55)	1.22(0.64)	0.17(0.14)	0.17(0.15)	43.85(37.24)	101.143(13.25)
KR-ABC (less)	0.15(0.13)	0.10(0.07)	1.11(0.12)	1.45(0.26)	0.23(0.41)	0.27(0.45)	67.57(47.11)	32.98(1.21)
K2-ABC	1.27(1.75)	0.20(0.23)	0.98(0.42)	1.46(1.10)	0.31(0.19)	0.61(0.82)	67.45(77.86)	23.47(1.59)
K-ABC	0.48(0.13)	0.14(0.06)	1.28(0.87)	1.42 (0.40)	0.22 (0.02)	0.27(0.25)	89.37(29.22)	30.66(2.57)
SMC-ABC (mean)	0.58 (0.15)	0.11(0.06)	1.03(0.46)	1.98(0.32)	0.28(0.15)	1.01(0.11)	170.41 (47.91)	38.50(2.34)
SMC-ABC (MAP)	0.51(0.28)	0.19(0.10)	0.89(0.33)	1.89(0.33)	0.53(0.46)	1.01(0.10)	163.19(42.51)	38.50(2.34)
ABC-DC	0.48(0.22)	0.25(0.13)	1.36(0.88)	1.55(0.12)	0.54(0.31)	1.17(0.11)	134.12(58.92)	29.94(4.57)
BO	0.83 (0.84)	0.16(0.24)	1.44(0.76)	1.09(0.50)	0.22(1.05)	0.45(0.41)	108.18(67.08)	3217.40(157.31)
MSM	0.65(0.16)	0.26(0.19)	1.50(0.59)	1.01(0.57)	0.14(0.13)	0.51(0.15)	89.17(33.20)	25.46(8.26)

Table 1 shows the separate errors made by each method for individual parameters. This was omitted from the main text due to space constraints.

C.2 Prior distribution for the parameters of the blowfly population dynamics

We describe here the prior distribution for the parameters $\theta := (P \in \mathbb{N},\ N_0 \in \mathbb{N},\ \sigma_d \in \mathbb{R}_+,\ \sigma_p \in \mathbb{R}_+,\ \tau \in \mathbb{N},\ \delta\ \in \mathbb{R}_+)$ in the blowfly population dynamics, the parameters that we used in our experiment. Let $\epsilon_p, \epsilon_{N_0}, \epsilon_{\sigma_d}, \epsilon_{\sigma_p}, \epsilon_\tau, \epsilon_\delta \sim \text{Normal}(0, 1)$ be independent standard Gaussian random variables. The prior can then be specified by defining the parameters as such random variables as

$$\begin{aligned} P &= \exp(2 + 2\epsilon_p), \\ N_0 &= \exp(5 + 0.5\epsilon_{N_0}), \\ \sigma_d &= \exp(-0.5 + \epsilon_{\sigma_d}), \\ \sigma_p &= \exp(-0.5 + \epsilon_{\sigma_p}), \\ \tau &= \exp(2 + \epsilon_\tau), \\ \delta &= \exp(-1 + 0.4\epsilon_\delta). \end{aligned}$$

Note that the parameters P, N_0, τ are to be rounded appropriately, as they are defined as being natural numbers.

D Supplementary materials for the experiments on alpha stable distributions in Section 4.4

D.1 Computation time

We offer here supplementary materials for the experiment on multivariate alpha stable distributions in Section 4.4 of the main text.

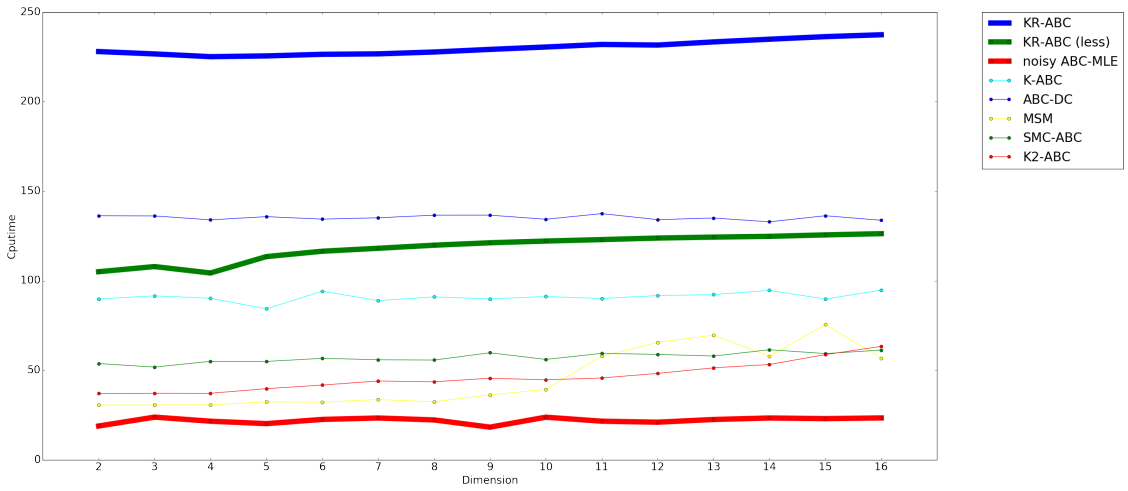

Figure 2: Computation time (in seconds) for the experiments in Section 4.4. We omit here the computation time of Bayesian optimization, but it was more than 1500 seconds for all the dimensions.

Figure D.1 shows computation time for each method in the experiments on multivariate alpha stable distributions in Section 4 of the main text, which information was omitted from the main body because of space constraints.

D.2 Definition of the deterministic map for sampling

We describe here the deterministic map τ_θ used in sampling multivariate alpha stable distributions (Chambers *et al.*, 1976), where $\theta := (\alpha, \beta, \mu, \sigma) \in (0, 2] \times [-1, 1] \times \mathbb{R} \times [0, \infty)$. Given $U_1 \sim \text{Unif}(-\pi/2, \pi/2)$ and $U_2 \sim \text{Exp}(1)$, the mapping $\tau_\theta(U_1, U_2) \in \mathbb{R}$ is defined as

$$\tau_\theta(U_1, U_2) := \sigma \tau_{\alpha,\beta}(U_1, U_2) + \mu,$$

where

$$\tau_{\alpha,\beta}(U_1, U_2) := \begin{Bmatrix} S_{\alpha,\beta} \frac{\sin[\alpha(U_1 + B_{\alpha,\beta})]}{[\cos(U_1)]^{1/\alpha}} \big(\frac{\cos[U_1 - \alpha(U_1 + B_{\alpha,\beta})]}{U_2} \big)^{(1-\alpha)/\alpha}, & \alpha \neq 1 \\ X = \frac{2}{\pi} [(\frac{\pi}{2} + \beta U_1) \tan U_1 - \beta \log(\frac{U_2 \cos U_1}{\frac{\pi}{2} + \beta U_1})], & \alpha = 1. \end{Bmatrix}$$

$$B_{\alpha,\beta} := \frac{\tan^{-1}(\beta \tan \frac{\pi\alpha}{2})}{\alpha}, \qquad S_{\alpha,\beta} := \left(1 + \beta^2 \tan^2 \frac{\pi\alpha}{2}\right)^{1/2\alpha}.$$

E Supplementary material for the pedestrian simulator experiment in Section 4.6

We present here supplementary material w.r.t. the pedestrian simulator experiment in Section 4.6.

E.1 Example of simulation results obtained with CrowdWalk

Figure 3: Example of simulation results obtained with Crowdwalk

Figure 3 shows an example of simulation results with the pedestrian flow simulator CrowdWalk (Yamashita *et al.*, 2010). The map is of Ginza, one of the largest commercial districts in Tokyo. Both green and red points indicate pedestrians, each of which is moving at an individual speed. Red points are pedestrians who are walking particularly slowly; these pedestrians are forced to walk slowly because the areas in which they are walking are crowded.

E.2 Parameters for the pedestrian simulator

Table 2: Parameters for the pedestrian simulator, including the fixed ones

Group	$\theta^{(S)}$	$\theta^{(G)}$	$\theta^{(P)}$	$\theta^{(R)}$	$\theta^{(N)}$	$\theta^{(T)}$
1	1	0	15, 29	5400, 1800	100	30
2	5	2	24, 43	5400, 5400	100	60
3	8	3	28, 48	5400, 5400	100	90
4	4	7	2, 0	1800, 3600	100	120
5	3	2	8, 9	5400, 5400	100	150
6	6	9	26, 14	5400, 3600	0	180
7	10	11	4, 41	1800, 3600	0	210
8	2	9	50, 18	1800, 3600	0	240
9	0	6	40, 33	3600, 5400	0	270
10	11	5	20, 25	1800, 3600	0	300

Table 2 shows the parameters of the 10 candidate groups in a mixture model used for parameter estimation. Note that the parameters $\theta^{(N)}$ and $\theta^{(T)}$ were *unknown* for each method since they were the parameters to be estimated. Groups 1 to 5 are the components of the true model, but this fact was also unknown for each method. The numbers in $\theta^{(S)}$, $\theta^{(G)}$ and $\theta^{(P)}$ indicate certain locations on the map (e.g., the Mitsukoshi Department Store, the Apple Store, and Ginza Station), which are predefined in terms of two-dimensional coordinates. The parameter $\theta^{(P)}$ indicates certain places where pedestrians in a single group visit. In this experiment, pedestrians in each group visited 2 intermediate places during the travel from the starting location to the goal; $\theta^{(R)}$ represent the respective durations of time (in seconds) at the intermediate places. (Note that the units for starting time $\theta^{(T)}$ are in minutes.)

E.3 Estimated parameters with the proposed method

Tables 3 and 4 show the estimated values for the parameters $\theta_i^{(N)}$ and $\theta_i^{(T)}$, respectively, for each of independent 20 trials. Recall that i in $\theta_i^{(N)}$ and $\theta_i^{(T)}$ is the index of 10 groups, i.e., $i = 1, \ldots, 10$. Results show that the proposed method was able to estimate the parameters of the 5 true groups in most cases. Note that the estimated values of $\theta_i^{(T)}$ for $i = 6, \ldots, 10$ were rather arbitrary. This is reasonable since the corresponding numbers of pedestrians $\theta_i^{(N)}$ in these groups were estimated to be zero or very small, and thus these groups could be treated as being nonexistent.

Table 3: Estimated values of $\theta_i^{(N)}$ with KR-ABC for each of 20 trials

Trial	$\theta_1^{(N)}$	$\theta_2^{(N)}$	$\theta_3^{(N)}$	$\theta_4^{(N)}$	$\theta_5^{(N)}$	$\theta_6^{(N)}$	$\theta_7^{(N)}$	$\theta_8^{(N)}$	$\theta_9^{(N)}$	$\theta_{10}^{(N)}$
1	95	132	102	1	79	32	2	9	13	31
2	100	105	105	84	100	0	0	0	0	0
3	98	102	90	81	101	0	16	0	7	0
4	93	98	99	101	95	0	0	0	0	9
5	103	12	84	9	88	27	2	33	19	3
6	87	105	108	100	92	0	6	0	0	0
7	90	102	104	102	97	0	0	0	0	0
8	116	79	93	118	88	0	0	0	0	1
9	97	91	101	110	94	0	0	1	0	0
10	102	127	0	0	97	0	38	70	51	12
11	102	105	94	87	100	0	0	0	0	7
12	0	2	100	228	103	0	0	0	1	64
13	98	97	96	108	95	0	0	0	1	0
14	103	98	94	94	102	3	0	0	0	1
15	105	176	106	0	9	78	2	14	3	2
16	96	134	100	0	95	0	70	0	0	0
17	98	106	96	87	98	4	1	4	0	2
18	798	101	97	97	98	0	2	1	1	0
19	109	54	55	181	68	0	0	31	0	0
20	98	90	101	102	99	3	2	0	0	0

Table 4: Estimated values of $\theta_i^{(T)}$ with KR-ABC for each of 20 trials

Trial	$\theta_1^{(T)}$	$\theta_2^{(T)}$	$\theta_3^{(T)}$	$\theta_4^{(T)}$	$\theta_5^{(T)}$	$\theta_6^{(T)}$	$\theta_7^{(T)}$	$\theta_8^{(T)}$	$\theta_9^{(T)}$	$\theta_{10}^{(T)}$
1	35	56	63	289	158	91	252	216	325	209
2	25	52	87	121	152	186	152	22	193	460
3	28	53	93	119	145	110	89	201	146	18
4	29	60	92	120	147	356	188	147	249	0
5	27	66	88	229	138	85	130	54	181	236
6	22	53	85	121	151	338	208	309	31	135
7	33	61	91	120	147	2	175	214	124	396
8	26	68	95	129	136	25	375	161	81	266
9	26	59	91	125	157	18	373	0	251	0
10	30	53	452	243	151	285	69	89	175	216
11	30	57	92	111	153	279	158	369	273	169
12	213	173	92	125	154	130	77	0	456	0
13	29	54	93	125	152	346	294	327	214	490
14	34	59	89	116	150	275	36	490	37	109
15	28	54	88	135	362	70	0	319	456	0
16	29	50	88	285	146	0	403	98	286	300
17	27	54	86	121	149	22	72	310	21	93
18	30	59	89	119	146	221	107	361	218	260
19	27	74	101	119	123	211	272	206	265	275
20	30	54	87	127	146	286	452	199	267	119

F Linear time estimator for the energy distance

In a way similar to that with Gretton *et al.* (2012, Section 6), we define here a linear-time estimator for the energy distance (Székely and Rizzo, 2013). Let $x_1 \ldots, x_n \sim P$ and $y_1, \ldots, y_n \sim Q$ be i.i.d. samples from the two distributions P and Q, and let $n_2 := \lfloor n/2 \rfloor$. The linear estimator can then be defined as

$$\frac{1}{n_2} \sum_{i=1}^{n_2} h((x_{2i-1}, y_{2i-1}), (x_{2i}, y_{2i})),$$

where

$$h((x_{2i-1}, y_{2i-1}), (x_{2i}, y_{2i})) := \|x_{2i-1} - y_{2i}\| + \|x_{2i} - y_{2i-1}\| - \|x_{2i-1} - x_{2i}\| - \|y_{2i-1} - y_{2i}\|.$$

It can be easily shown that this is unbiased and converges to the population energy distance between P and Q at a rate of $O_p(n^{-1/2})$, as Gretton *et al.* (2012, Theorem 15) showed for a linear estimator in MMD. The above linear estimator can be computed at a cost of $O(n)$, which is less than the cost of $O(n^2)$ required for an ordinary quadratic estimator. (Note, however, that the linear estimator has higher variance than a quadratic one.)

References

Billingsley, P. (1999). *Convergence of Probability Measures, 2nd Edition*. Wiley-Interscience.

Chambers, J. M., Mallows, C. L., and Stuck, B. (1976). A method for simulating stable random variables. *Journal of the american statistical association*, **71**(354), 340–344.

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

Lele, S. R., Nadeem, K., , and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, **105**(492), 1617–1625.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, **143**(8), 1249–1272.

Yamashita, T., Soeda, S., and Noda, I. (2010). Assistance of evacuation planning with high-speed network model-based pedestrian simulator. In *Proceedings of Fifth International Conference on Pedestrian and Evacuation Dynamics (PED 2010)*, page 58. PED 2010.