K2-ABC: Approximate Bayesian Computation with Kernel Embeddings

Mijung Park[1] Wittawat Jitkrittum[1] Dino Sejdinovic[2]

Gatsby Unit, University College London[1] University of Oxford[2]

Summary

- ABC = Bayesian inference paradigm: likelihood is intractable but generating (pseudo) data given any parameter value is easy.
- Approximate posterior defined through a *measure of similarity* between pseudo data and the observed data.
- Similarity typically measured through summary statistics: difficult to define, requires domain experts and often introduces a difficult-to-quantify information loss when statistics are not sufficient.
- **Contribution**: Use Maximum Mean Discrepancy (MMD), a kernel-based distance between probability measures, to define this similarity.
- No information loss for a broad family of kernels.

Approximate Bayesian Computation (ABC)

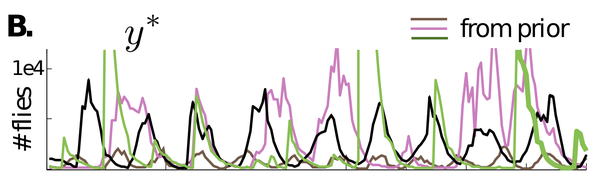
- **Given**: Prior $p(\theta)$, intractable likelihood $p(\mathbf{Y}|\theta)$, observed set $\mathbf{Y}$.
- **Goal**: Sample from $p(\theta|\mathbf{Y}) \propto p(\theta)p(\mathbf{Y}|\theta)$.
- **Problem**: Cannot evaluate $p(\mathbf{Y}|\theta)$. Can sample $\mathbf{X} \sim p(\cdot|\theta)$ easily.

Example: a complicated dynamical system for blow fly population

$$N_{t+1} = PN_{t-\tau}\exp\left(-\frac{N_{t-\tau}}{N_0}\right)e_t + N_t\exp(-\delta\varepsilon_t)$$

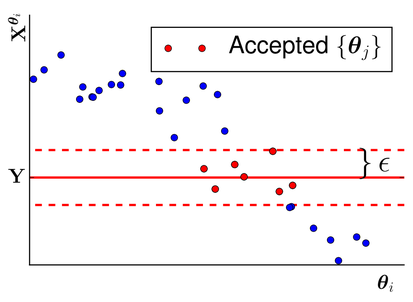
where $e_t \sim \text{Gamma}\left(\frac{1}{\sigma_P^2}, \sigma_P^2\right)$ and $\varepsilon_t \sim \text{Gamma}\left(\frac{1}{\sigma_d^2}, \sigma_d^2\right)$.

- $\theta := \{P, N_0, \sigma_d, \sigma_p, \tau, \delta\}$
- Given $\mathbf{Y} = \{N_1, \ldots, N_T\}$, want to sample from $p(\theta|\mathbf{Y})$.



ABC Likelihood $p_\varepsilon(\mathbf{Y}|\theta)$

$$\begin{aligned} p(\mathbf{Y}|\theta) &= \int p(\mathbf{X}|\theta)\delta(\mathbf{X}-\mathbf{Y})\,\mathrm{d}\mathbf{X} \\ &\approx \int p(\mathbf{X}|\theta)\kappa_\varepsilon(\mathbf{X},\mathbf{Y})\,\mathrm{d}\mathbf{X} := p_\varepsilon(\mathbf{Y}|\theta) \\ &\approx \kappa_\varepsilon(\mathbf{X}^\theta, \mathbf{Y}) \text{ where } \mathbf{X}^\theta \sim p(\cdot|\theta), \end{aligned}$$



- $\kappa_\varepsilon(\mathbf{X}, \mathbf{Y})$ defines **similarity** between $\mathbf{X}$ and $\mathbf{Y}$.
- ABC algorithms sample from $p_\varepsilon(\theta|\mathbf{Y}) \propto p(\theta)p_\varepsilon(\mathbf{Y}|\theta)$
- Commonly used rejection ABC sets $\kappa_\varepsilon(\mathbf{X}, \mathbf{Y}) := \mathbf{1}[\|s(\mathbf{X}) - s(\mathbf{Y})\|_2 < \varepsilon]$.
  - s : function to compute summary statistics
  - $\mathbf{1}[\cdot] \in \{0, 1\}$: indicator function

Problems with Summary Statistics $s(\cdot)$

- Difficult to design **sufficient statistics**.
- More statistics give high sufficiency. But, higher rejection rate.
- Insufficient $s(\cdot)$ will lead to an incorrect posterior.

Contribution: Kernel-Based Similarity

- Use a kernel distance MMD to define similarity κ_ε. No need for $s(\cdot)$.

Rejection ABC

$$\kappa_\varepsilon(\mathbf{X}, \mathbf{Y}) = \mathbf{1}[\|s(\mathbf{X}) - s(\mathbf{Y})\|_2 < \varepsilon]$$

K2-ABC (Proposed)

$$\kappa_\varepsilon(\mathbf{X}, \mathbf{Y}) = \exp\left(-\widehat{\text{MMD}}(\mathbf{X}, \mathbf{Y})^2/\varepsilon\right)$$

- MMD can detect any difference in distributions without the need of handcrafted summary statistics.

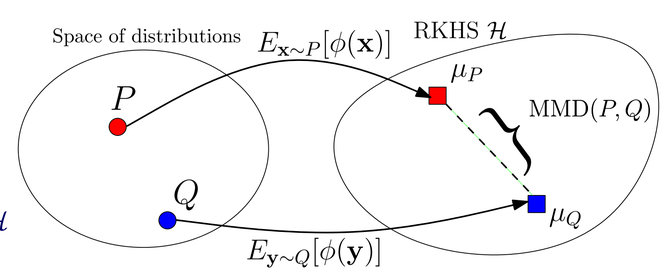
Maximum Mean Discrepancy (MMD) [1]

$$\begin{aligned} \text{MMD}(P, Q) &= \sup_{\|f\|_{\mathcal{H}} \le 1} \mathbb{E}_{\mathbf{x}\sim P}[f(\mathbf{x})] - \mathbb{E}_{\mathbf{y}\sim Q}[f(\mathbf{y})] = \|\mu_P - \mu_Q\|_{\mathcal{H}} \\ &\approx \left[\frac{1}{n(n-1)}\sum_{i\neq j} k(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n(n-1)}\sum_{i\neq j} k(\mathbf{y}_i, \mathbf{y}_j) - \frac{2}{n^2}\sum_{i,j=1}^{n} k(\mathbf{x}_i, \mathbf{y}_j)\right]^{1/2} \end{aligned}$$

- Nonparametric distance

$$\text{MMD}(P, Q) = 0 \iff P = Q$$

- Kernel $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{H}}$



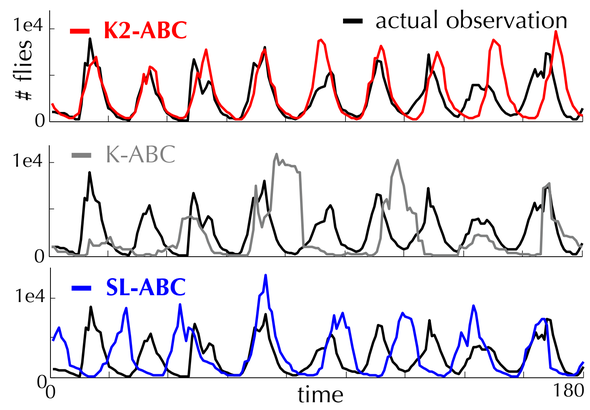
- $\mu_{p(\cdot|\theta)}$ is always sufficient.
- Intuitively, $\mu_{p(\cdot|\theta)}$ contains all moments of $p(\cdot|\theta)$.

K2-ABC (Proposed)

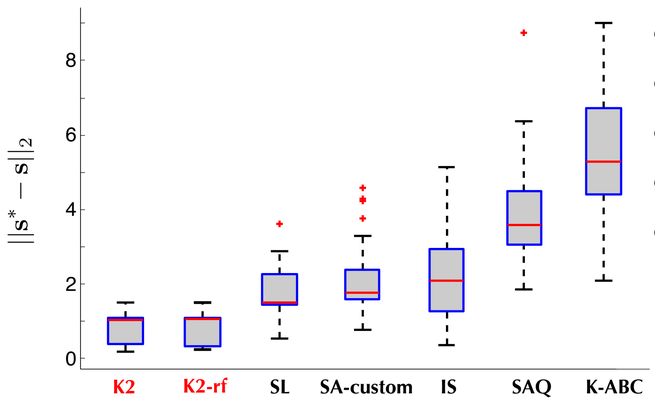
Output: Posterior empirical distribution $\sum_{i=1}^M w_i \delta_{\theta_i}$

1: **for** $i = 1, \ldots, M$ **do**
2: Sample $\theta_i \sim p(\theta)$
3: Sample pseudo dataset $\mathbf{X}_i \sim p(\cdot|\theta_i)$
4: $\widetilde{w}_i = \kappa_\varepsilon(\mathbf{X}_i, \mathbf{Y}) = \exp\left(-\widehat{\text{MMD}}(\mathbf{X}_i, \mathbf{Y})^2/\varepsilon\right)$
5: Set $w_i = \widetilde{w}_i / \sum_{j=1}^M \widetilde{w}_j$ for $i = 1, \ldots, M$

Experiments on Blow Fly Data



- Simulated trajectories with inferred posterior mean of θ.
- Other methods use handcrafted 10-dim. summary statistics (as in [2, 3]) of the marginal distributions.



- $\tilde{\theta}$:= posterior mean.
- Simulate $\mathbf{X} \sim p(\cdot|\tilde{\theta})$ 100 times.
- $\mathbf{s} = s(\mathbf{X})$ and $\mathbf{s}^* = s(\mathbf{Y})$.
- K2-ABC yields lowest error on $\mathbf{s}$.
- Linear-time K2-ABC with random features (K2-rf) has an indistinguishable performance.

K2-ABC can infer correct θ without the need for handcrafted $s(\cdot)$.

References

1. A. Gretton et. al. A kernel two-sample test. JMLR 2012.
2. E. Meeds & M. Welling. GPS-ABC: Gaussian process surrogate approximate Bayesian computation. UAI 2014.
3. S. N. Wood. Statistical inference for noisy nonlinear ecological dynamic systems. Nature 2010.

MP and WJ thank the Gatsby Charitable Foundation for the financial support.

Contact: wittawat@gatsby.ucl.ac.uk

Code: github.com/wittawatj/k2abc