K2-ABC: Approximate Bayesian Computation with Kernel Embeddings

Mijung Park*,[1] **Wittawat Jitkrittum***,[1] Dino Sejdinovic†

*Gatsby Unit, University College London
†University of Oxford

AISTATS 2016, Cadiz, Spain

9 May 2016

[1] MP and WJ contributed equally.

Approximate Bayesian Computation (ABC)

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

Example: a complicated dynamical system for blow fly population

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

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

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

Approximate Bayesian Computation (ABC)

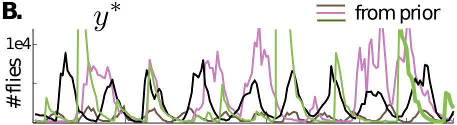
- Given: prior $p(\boldsymbol{\theta})$, intractable likelihood $p(\mathbf{Y}|\boldsymbol{\theta})$, observations $\mathbf{Y}$.
- Goal: Sample from $p(\boldsymbol{\theta}|\mathbf{Y}) \propto p(\boldsymbol{\theta})p(\mathbf{Y}|\boldsymbol{\theta})$.
- Problem: Cannot evaluate $p(\mathbf{Y}|\boldsymbol{\theta})$.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

Example: a complicated dynamical system for blow fly population

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

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

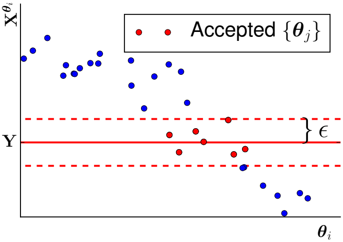
- $\boldsymbol{\theta} := \{P, N_0, \sigma_d, \sigma_p, \tau, \delta\}$
- Given $\mathbf{Y} = \{N_1, \ldots, N_T\}$, want to sample from $p(\boldsymbol{\theta}|\mathbf{Y})$.



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

- Observe a dataset $\mathbf{Y}$,

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

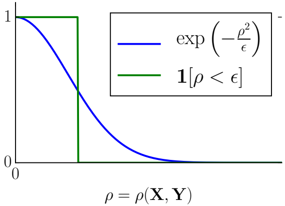


where $\kappa_\epsilon(\mathbf{X},\mathbf{Y})$ defines similarity between $\mathbf{X}$ and $\mathbf{Y}$.

- Commonly used **rejection ABC** sets

$$\kappa_\epsilon(\mathbf{X},\mathbf{Y}) := \mathbf{1}[\rho(\mathbf{X},\mathbf{Y}) < \epsilon],$$

- Distance $\rho(\mathbf{X},\mathbf{Y}) := \|s(\mathbf{X}) - s(\mathbf{Y})\|_2$
- $\mathbf{1}[\cdot] \in \{0,1\}$: indicator function
- s : function to compute summary statistics



Summary Statistics $s(\cdot)$

- Difficult to choose summary statistics $s(\cdot)$ in

$$\rho(\mathbf{X}, \mathbf{Y}) = \|s(\mathbf{X}) - s(\mathbf{Y})\|_2.$$

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

Contribution:

- Use a kernel distance MMD to define ρ. No need to design $s(\cdot)$.

Rejection ABC:	K2-ABC (proposed):
$\rho(\mathbf{X}, \mathbf{Y}) = \|s(\mathbf{X}) - s(\mathbf{Y})\|_2$	$\rho(\mathbf{X}, \mathbf{Y}) = \widehat{\mathrm{MMD}}(\mathbf{X}, \mathbf{Y})$

Summary Statistics $s(\cdot)$

- Difficult to choose summary statistics $s(\cdot)$ in

$$\rho(\mathbf{X}, \mathbf{Y}) = \|s(\mathbf{X}) - s(\mathbf{Y})\|_2.$$

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

Contribution:

- Use a kernel distance MMD to define ρ. No need to design $s(\cdot)$.

Rejection ABC:

$$\rho(\mathbf{X}, \mathbf{Y}) = \|s(\mathbf{X}) - s(\mathbf{Y})\|_2$$

K2-ABC (proposed):

$$\rho(\mathbf{X}, \mathbf{Y}) = \widehat{\mathrm{MMD}}(\mathbf{X}, \mathbf{Y})$$

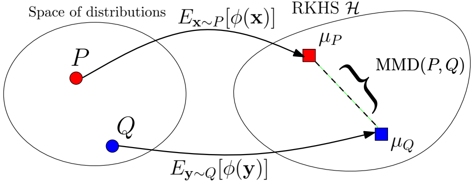
Maximum Mean Discrepancy (MMD) [Gretton et al., 2006]

$$\mathrm{MMD}^2(P,Q) = \|\mathbb{E}_{\mathbf{x}\sim P}[\phi(\mathbf{x})] - \mathbb{E}_{\mathbf{y}\sim Q}[\phi(\mathbf{y})]\|_{\mathcal{H}}^2 \approx \widehat{\mathrm{MMD}}^2(\mathbf{X},\mathbf{Y})$$

$$:= \frac{1}{n(n-1)}\sum_{i\neq j} k(\mathbf{x}_i,\mathbf{x}_j) + \frac{1}{n(n-1)}\sum_{i\neq j} k(\mathbf{y}_i,\mathbf{y}_j) - \frac{2}{n^2}\sum_{i,j=1}^{n} k(\mathbf{x}_i,\mathbf{y}_j)$$

- If kernel k is characteristic (e.g., Gaussian kernel), $\mu_P = \mathbb{E}_{\mathbf{x}\sim P}[\phi(\mathbf{x})]$ is sufficient for P.

- $k(\mathbf{x},\mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y})\rangle_{\mathcal{H}}$
- Intuitively, μ_P contains all moments of P.



K2-ABC (Proposed Method)

- To sample $\{\boldsymbol{\theta}_i\}_{i=1}^M \sim p_\epsilon(\boldsymbol{\theta}|\mathbf{Y})$, do

Output: Approximate posterior $\sum_{i=1}^M \delta_{\boldsymbol{\theta}_i} w_i$

```
for i = 1, ..., M do
    Sample θ_i ~ p(θ)
    Sample pseudo dataset X_i ~ p(·|θ_i)
    w̃_i = κ_ε(X_i, Y) = exp( − MMD^2(X_i, Y) / ε )
end for
w_i = w̃_i / Σ_{j=1}^M w̃_j for i = 1, ..., M
return {θ_i}_{i=1}^M with weights {w_i}_{i=1}^M
```

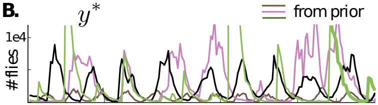
for $i = 1, \ldots, M$ **do**
 Sample $\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta})$
 Sample pseudo dataset $\mathbf{X}_i \sim p(\cdot|\boldsymbol{\theta}_i)$
 $\widetilde{w}_i = \kappa_\epsilon(\mathbf{X}_i, \mathbf{Y}) = \exp\left(-\frac{\widehat{\text{MMD}}^2(\mathbf{X}_i, \mathbf{Y})}{\epsilon}\right)$
end for
$w_i = \widetilde{w}_i / \sum_{j=1}^M \widetilde{w}_j$ for $i = 1, \ldots, M$
return $\{\boldsymbol{\theta_i}\}_{i=1}^M$ with weights $\{w_i\}_{i=1}^M$

- Given a function g,

$$\mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathbf{Y})}[g(\boldsymbol{\theta})] \approx \frac{1}{M} \sum_{i=1}^M w_i g(\boldsymbol{\theta_i}).$$

Blow Fly Population Modelling

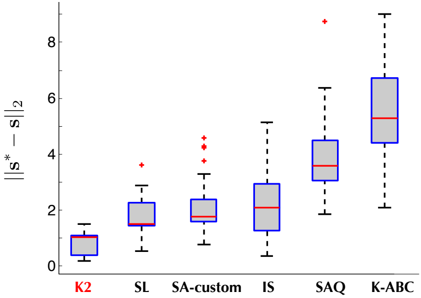
$$N_{t+1} = PN_{t-\tau} \exp\left(-\frac{N_{t-\tau}}{N_0}\right) e_t + N_t \exp(-\delta\epsilon_t)$$



- $e_t \sim$ Gam $\left(\frac{1}{\sigma_P^2}, \sigma_P^2\right)$ and $\epsilon_t \sim$ Gam $\left(\frac{1}{\sigma_d^2}, \sigma_d^2\right)$.
- Observe $\mathbf{Y}$ (black solid line).
- Want posterior of $\boldsymbol{\theta} := \{P, N_0, \sigma_d, \sigma_p, \tau, \delta\}$.

- Compare 6 ABC methods.
- 5 other methods use handcrafted 10-dim. summary statistics [Meeds and Welling, 2014].
  - quantiles of the marginal distribution
  - quantiles of first-order differences
  - maximal peaks

Errors on Summary Statistics



- $\tilde{\boldsymbol{\theta}} :=$ posterior mean.
- Simulate $\mathbf{X} \sim p(\cdot|\tilde{\boldsymbol{\theta}})$ 100 times.
- $\mathbf{s} = s(\mathbf{X})$ and $\mathbf{s}^* = s(\mathbf{Y})$.

- $\tilde{\boldsymbol{\theta}}$ inferred by K2-ABC gives lowest error on $\mathbf{s}$.
- Recall that K2-ABC does not use $\mathbf{s}$, unlike others.

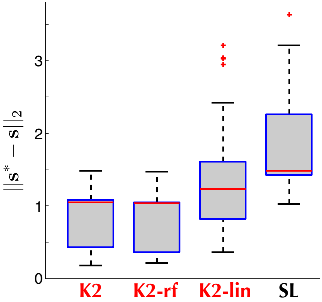
K2-ABC can infer the generative parameters without the need for handcrafted summary statistics.

Linear-Time K2-ABC

- $\widehat{\mathrm{MMD}}^2(\mathbf{X}, \mathbf{Y})$ costs $O(n^2)$ where $n =$ sample size. Expensive.

Solutions:

1. Linear-time unbiased estimator. Costs $O(n)$.
2. Random Fourier features $\hat{\phi}(\mathbf{x}) \in \mathbb{R}^D$ such that $k(\mathbf{x}, \mathbf{y}) \approx \hat{\phi}(\mathbf{x})^\top \hat{\phi}(\mathbf{y})$. Costs $O(Dn)$. We set $D = 50$.



K2-ABC with random features performs equally well with a much cheaper cost.

Summary

ABC problem:

- Goal: Sample from $p(\boldsymbol{\theta}|\mathbf{Y})$ where the likelihood is intractable.
- Can only sample from the likelihood.

Solution:

- Idea: Keep $\boldsymbol{\theta}$ such that $\mathbf{X} \sim p(\cdot|\boldsymbol{\theta})$ is "*similar*" to $\mathbf{Y}$.
- **Contribution**: K2-ABC uses kernel MMD to define the similarity.
  - No need to design summary statistics.
  - Capture all information of $p(\cdot|\boldsymbol{\theta})$.
- Code: `https://github.com/wittawatj/k2abc`

Tue May 10. Poster 6.

Questions?

Thank you

References I

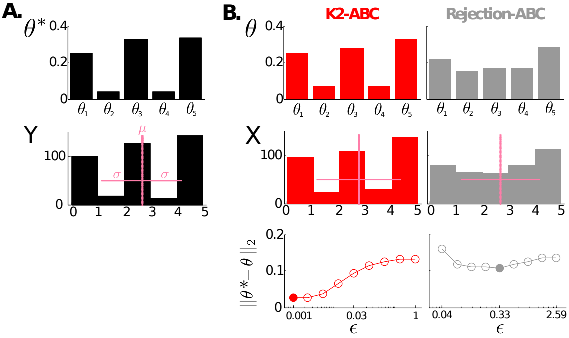
- K2-ABC on Arxiv: http://arxiv.org/abs/1502.02558

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. J. (2006).
A kernel method for the two-sample-problem.
In *Advances in neural information processing systems*, pages 513–520.

Meeds, E. and Welling, M. (2014).
GPS-ABC: Gaussian Process Surrogate Approximate Bayesian Computation.
In *UAI*, volume 30, pages 593–601.

Toy Problem: Failure of Insufficient Statistics

$$
\begin{aligned}
p(y|\boldsymbol{\theta}) &= \sum_{i=1}^{5} \theta_i \text{Uniform}(y; [i-1, i]) \\
\pi(\boldsymbol{\theta}) &= \text{Dirichlet}(\boldsymbol{\theta}; \mathbf{1}) \\
\boldsymbol{\theta}^* &= \text{(see figure A)}
\end{aligned}
$$



- $s(\mathbf{X}) = (\hat{\mathbb{E}}[x], \hat{\mathbb{V}}[x])^\top$ for Rejection and Soft ABC.
- Insufficient to represent $p(y|\theta)$.

Rejection ABC Algorithm

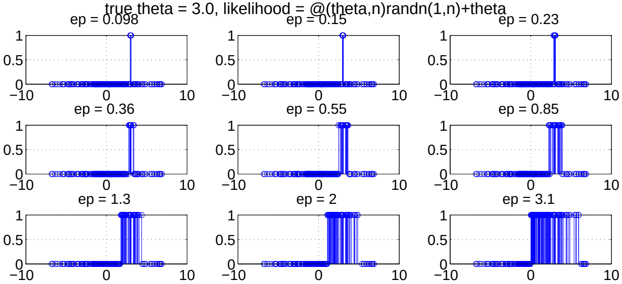
- **Input:** observed dataset $\mathbf{Y}$, distance ρ, threshold ϵ
- **Output:** posterior sample $\{\boldsymbol{\theta}_i\}_{i=1}^{M}$ from approximate posterior $p_\epsilon(\boldsymbol{\theta}|\mathbf{Y}) \propto p(\boldsymbol{\theta})p_\epsilon(\mathbf{Y}|\boldsymbol{\theta})$

```
repeat
    Sample θ ~ p(θ)
    Sample a pseudo dataset X ~ p(·|θ)
    if ρ(X, Y) < ε then
        Keep θ
    end if
until we have M points
```

- **Notation**: $\mathbf{Y} =$ observed set. $\mathbf{X} =$ pseudo (generated) dataset.

Rejection ABC Example

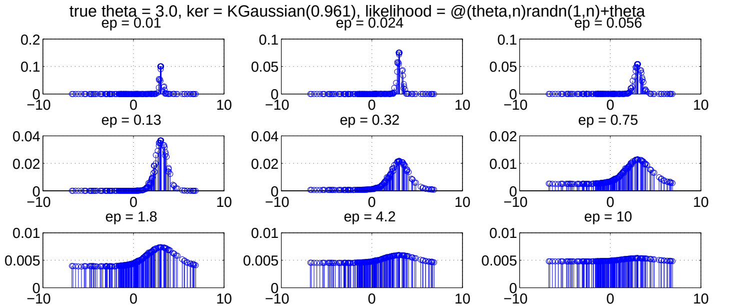
$$
\begin{aligned}
p(y|\theta) &= \mathcal{N}(y; \theta, 1) \\
p(\theta) &= \mathcal{N}(\theta, 0, 8) \\
\theta^* &= 3.0 \\
\rho(\mathbf{X}, \mathbf{Y}) &= \left| \hat{\mathbb{E}}_{\mathbf{X}}[x] - \hat{\mathbb{E}}_{\mathbf{Y}}[y] \right|
\end{aligned}
$$



- Low $\epsilon \Rightarrow$ sample closely follows true posterior. High rejection rate.
- High $\epsilon \Rightarrow$ get θ sample from prior.

1D Gaussian Example with K2-ABC

$$
\begin{aligned}
p(y|\theta) &= \mathcal{N}(y;\theta,1) \\
\pi(\theta) &= \mathcal{N}(\theta,0,8) \\
\theta^* &= 3.0
\end{aligned}
$$



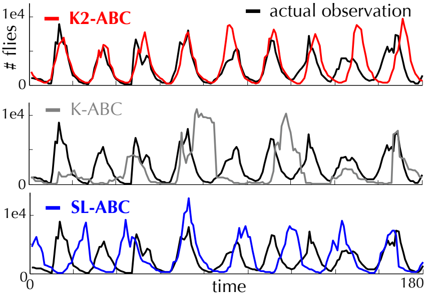
- High $\epsilon \Rightarrow$ get θ sample from prior
- Low $\epsilon \Rightarrow$ sample closely follows true posterior.

Simulated Trajectories

Number of blow flies over time

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

- $e_t \sim$ Gam $\left(\frac{1}{\sigma_P^2}, \sigma_P^2\right)$ and $\epsilon_t \sim$ Gam $\left(\frac{1}{\sigma_d^2}, \sigma_d^2\right)$.
- Want $\boldsymbol{\theta} := \{P, N_0, \sigma_d, \sigma_p, \tau, \delta\}$.



- ← Simulated trajectories with inferred posterior mean of $\boldsymbol{\theta}$
- Other methods use handcrafted 10-dim. summary statistics [Meeds and Welling, 2014].
  - quantiles of the marginal distribution
  - quantiles of first-order differences
  - maximal peaks