# Bayesian Indirect Inference Using a Parametric Auxiliary Model

Wittawat Jitkrittum
wittawat@gatsby.ucl.ac.uk

Gatsby Machine Learning Journal Club

29 Feb 2016

# Approximate Bayesian Computation (ABC)

- Given a tractable prior $p(\boldsymbol{\theta})$, an intractable likelihood $p(\mathbf{y}|\boldsymbol{\theta})$.
- Observe a set $\mathbf{y}$.
- **Goal**: get sample from posterior $p(\boldsymbol{\theta}|\mathbf{y})$.
- Possible to sample $\mathbf{x} \sim p(\cdot|\boldsymbol{\theta})$ easily.

**Example**: a complicated dynamical system for blowfly 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 $\left(\frac{1}{\sigma_d^2}, \sigma_d^2\right)$.

- $\boldsymbol{\theta} := \{P, N_0, \sigma_d, \sigma_p, \tau, \delta\}$

**Basic idea of ABC**:

- Find $\boldsymbol{\theta}$'s such that $\mathbf{x} \sim p(\cdot|\boldsymbol{\theta})$ is close to $\mathbf{y}$.

# Rejection ABC

- The most basic form of ABC.

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{y}) &\propto p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta}) \\
&= p(\boldsymbol{\theta}) \int p(\mathbf{x}|\boldsymbol{\theta})\, \mathrm{d}\delta_{\mathbf{y}}(\mathbf{x}) \\
&\approx p(\boldsymbol{\theta}) \int I(\kappa(\mathbf{x}, \mathbf{y}) < \epsilon) p(\mathbf{x}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{x},
\end{aligned}
$$

where $\kappa(\mathbf{x}, \mathbf{y})$ is low when $\mathbf{x}$ (pseudo-dataset) is close to $\mathbf{y}$.

```
1: repeat
2:    Sample θ_i ~ p(·)
3:    Sample a dataset x_i ~ p(·|θ_i)
4:    if κ(x_i, y) < ε then
5:       Keep θ_i
6:    end if
7: until we have enough samples
8: return posterior sample {θ_i}_i
```

# ABC Likelihood and Summary Statistics

$$p(\boldsymbol{\theta}|\mathbf{y}) \approx p(\boldsymbol{\theta}) \int I(\kappa(\mathbf{x},\mathbf{y}) < \epsilon) p(\mathbf{x}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{x}$$

- Typically, $\kappa(\mathbf{x},\mathbf{y}) := \rho(s(\mathbf{x}), s(\mathbf{y}))$
  - $s(\mathbf{x})$ = summary statistics of $\mathbf{x}$.
  - $\rho(s(\mathbf{x}), s(\mathbf{y}))$ = distance on summary statistics
- Define a kernel weighting function $K_\epsilon(t)$ s.t. $K_\epsilon(t)$ high around 0,
- ABC likelihood:

$$p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) = \int K_\epsilon(\rho(s(\mathbf{x}), s(\mathbf{y}))) p(\mathbf{x}|\boldsymbol{\theta})\, \mathrm{d}\mathbf{x}.$$

- $K_\epsilon(t) = I(t < \epsilon) \in \{0, 1\}$ is a kind of weighting function.
- If $s(\cdot)$ is sufficient, then $p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \to p(\boldsymbol{\theta}|\mathbf{y})$ as $\epsilon \to 0$ [Blum et al., 2013].

# ABC Likelihood and Summary Statistics

$$p(\boldsymbol{\theta}|\mathbf{y}) \approx p(\boldsymbol{\theta}) \int I(\kappa(\mathbf{x},\mathbf{y}) < \epsilon) p(\mathbf{x}|\boldsymbol{\theta}) \, \mathrm{d}\mathbf{x}$$

- Typically, $\kappa(\mathbf{x},\mathbf{y}) := \rho(s(\mathbf{x}), s(\mathbf{y}))$
  - $s(\mathbf{x})$ = summary statistics of $\mathbf{x}$.
  - $\rho(s(\mathbf{x}), s(\mathbf{y}))$ = distance on summary statistics
- Define a kernel weighting function $K_\epsilon(t)$ s.t. $K_\epsilon(t)$ high around 0,
- ABC likelihood:

$$p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) = \int K_\epsilon(\rho(s(\mathbf{x}), s(\mathbf{y}))) p(\mathbf{x}|\boldsymbol{\theta}) \, \mathrm{d}\mathbf{x}.$$

- $K_\epsilon(t) = I(t < \epsilon) \in \{0, 1\}$ is a kind of weighting function.
- If $s(\cdot)$ is sufficient, then $p_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \to p(\boldsymbol{\theta}|\mathbf{y})$ as $\epsilon \to 0$ [Blum et al., 2013].

# MCMC ABC [Marjoram et al., 2003]

- $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ : MCMC proposal distribution.
- $T$ : number of iterations

**Algorithm 1** MCMC ABC algorithm of Marjoram et al. (2003).

Set $\boldsymbol{\theta}^0$
Simulate $\mathbf{x}^0 \sim p(\cdot|\boldsymbol{\theta}^0)$
**for** $i = 1$ **to** $T$ **do**
  Draw $\boldsymbol{\theta}^* \sim q(\cdot|\boldsymbol{\theta}^{i-1})$
  Simulate $\mathbf{x}^* \sim p(\cdot|\boldsymbol{\theta}^*)$
  Compute $r = \min(1, \frac{p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{i-1}|\boldsymbol{\theta}^*)K_\varepsilon(\rho(s(\mathbf{x}^*), s(\mathbf{y})))}{p(\boldsymbol{\theta}^{i-1})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1})K_\varepsilon(\rho(s(\mathbf{x}^{i-1}), s(\mathbf{y})))})$
  **if** $\mathrm{uniform}(0,1) < r$ **then**
    $\boldsymbol{\theta}^i = \boldsymbol{\theta}^*$ and $\mathbf{x}^i = \mathbf{x}^*$
  **else**
    $\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1}$ and $\mathbf{x}^i = \mathbf{x}^{i-1}$
  **end if**
**end for**

- More computationally efficient than rejection ABC (on a single machine).

# This Paper

Bayesian Indirect Inference Using a Parametric Auxiliary Model
Christopher C. Drovandi, Anthony N. Pettitt, Anthony Lee
http://arxiv.org/abs/1505.03372

- Overview of a class of ABC algorithms that uses an auxiliary model.
- Comment: rejection ABC does not use an auxiliary model.
- Tractable auxiliary model $p_A(\mathbf{y}|\boldsymbol{\phi})$ in place of $p(\mathbf{y}|\boldsymbol{\theta})$.
- "**Indirect inference**"
- Summary statistics can be formulated based on $p_A$.
- Will go through many such algorithms.

# ABC Indirect Parameter (IP) [Drovandi et al., 2011]

- Summary statistic = parameter estimate of the auxiliary model.
- For each simulated $\mathbf{x} \sim p(\cdot|\boldsymbol{\theta})$, maximum likelihood estimate

$\arg\max_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} p_A(\mathbf{x}|\boldsymbol{\phi})$ defines a noisy map $\boldsymbol{\theta} \mapsto \boldsymbol{\phi}(\mathbf{x})$

- Set $s(\mathbf{x}) := \boldsymbol{\phi}(\mathbf{x})$.
- Mahalanobis distance:

$$\rho(s(\mathbf{x}), s(\mathbf{y})) = \sqrt{(\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{y}))^\top \mathbf{J}(\boldsymbol{\phi}(\mathbf{y}))(\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{y}))},$$

where

$$\mathbf{J}(\boldsymbol{\phi}(\mathbf{y})) = \frac{1}{|\mathbf{y}|}\sum_{y\in\mathbf{y}} [\nabla \log p_A(y|\boldsymbol{\phi}(\mathbf{y}))][\nabla \log p_A(y|\boldsymbol{\phi}(\mathbf{y}))]^\top$$

is the empirical Fisher information matrix of $p_A$.

# Comments on ABC IP

$$\rho(s(\mathbf{x}), s(\mathbf{y})) = \sqrt{(\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{y}))^\top \mathbf{J}(\boldsymbol{\phi}(\mathbf{y}))(\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{y}))},$$

- More efficient than

$$\rho(s(\mathbf{x}), s(\mathbf{y})) = \|\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{y})\|_2.$$

  $\mathbf{J}(\boldsymbol{\phi}(\mathbf{y}))$ takes into account the variance.
- Useful if $p_A$ fits $\mathbf{y}$ well.
- **Assumption 1**: No non-identifiability issue i.e., $\arg\max_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} p_A(\mathbf{x}|\boldsymbol{\phi})$ is unique for all $\boldsymbol{\theta}$ with positive prior support.
- Otherwise, possible that $\rho(s(\mathbf{x}), s(\mathbf{y})) > 0$ when $\mathbf{x} = \mathbf{y}$.

# ABC Indirect Likelihood (IL) [Gleim and Pigorsch, 2013]

$$\rho(s(\mathbf{x}), s(\mathbf{y})) = \log p_A(\mathbf{y}|\boldsymbol{\phi}(\mathbf{y})) - \log p_A(\mathbf{y}|\boldsymbol{\phi}(\mathbf{x}))$$

- Recall $\boldsymbol{\phi}(\mathbf{x}) = \arg\max_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} p_A(\mathbf{x}|\boldsymbol{\phi})$.
- $\log p_A(\mathbf{y}|\boldsymbol{\phi}(\mathbf{y}))$ is fixed.
- $\log p_A(\mathbf{y}|\boldsymbol{\phi}(\mathbf{y})) \geq \log p_A(\mathbf{y}|\boldsymbol{\phi}(\mathbf{x}))$ for all simulated $\mathbf{x}$.
- Same summary statistics as ABC IP. Use likelihood-based discrepancy.
- **Assumption 2**: $p_A(\mathbf{y}|\phi(\mathbf{x}, \boldsymbol{\theta}))$ is unique for all $\boldsymbol{\theta}$ with positive prior support.
- Not require differentiability of $\log p_A$.

# ABC Indirect Score (IS) [Gleim and Pigorsch, 2013]

- Score vector of auxiliary model

$$\mathbf{S}_A(\mathbf{y}, \boldsymbol{\phi}) = \left( \frac{\partial \log p_A(\mathbf{y}|\boldsymbol{\phi})}{\partial \phi_1}, \ldots, \frac{\partial \log p_A(\mathbf{y}|\boldsymbol{\phi})}{\partial \phi_{\dim(\boldsymbol{\phi})}} \right)$$

- **Known**: $\mathbf{S}_A(\mathbf{y}, \boldsymbol{\phi}(\mathbf{y})) = 0$ i.e., score evaluated at MLE.
- **Idea**: Search for $\boldsymbol{\theta}$ that leads to $\mathbf{x}$, that produces $\mathbf{S}_A(\mathbf{x}, \boldsymbol{\phi}(\mathbf{y})) \approx 0$.
- Discrepancy:

$$\rho(s(\mathbf{x}), s(\mathbf{y})) = \sqrt{\mathbf{S}_A(\mathbf{x}, \boldsymbol{\phi}(\mathbf{y}))^\top \mathbf{J}(\boldsymbol{\phi}(\mathbf{y}))^{-1} \mathbf{S}_A(\mathbf{x}, \boldsymbol{\phi}(\mathbf{y}))}.$$

- No need to estimate MLE for each simulated $\mathbf{x}$. Computationally cheap.
- **Assumption 3**: $\mathbf{S}_A(\mathbf{x}, \boldsymbol{\phi}(\mathbf{y}))$ is unique for all $\mathbf{x}$.

# Parametric Bayesian Indirect Likelihood (pdBIL)

[Reeves and Pettitt, 2005, Gallant and McCulloch, 2009]

- $n$ replicate simulated datasets
- Artificial likelihood:

$$p_{A,n}(\mathbf{y}|\boldsymbol{\theta}) = \int p_A(\mathbf{y}|\boldsymbol{\phi}_n(\boldsymbol{\theta}, \mathbf{x}_{1:n})) \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta}) \, \mathrm{d}\mathbf{x}_{1:n}.$$

- Just another (stochastic) likelihood. Can use any Bayesian algorithm e.g., MCMC.
- Related to ABC IL. But no discrepancy $\rho$.
- No comparison of summary stats $\implies$ No need $\epsilon$.

# Example: True Posterior under pdBIL

- If $p$ is contained in $p_A$, pdBIL will target the true posterior as $n \to \infty$.
- True model: $p(y|\theta) = \mathcal{N}(y; \theta, 1)$
- Auxiliary model (Gaussian mixture):

$$p_A(y|\boldsymbol{\theta}) = w\mathcal{N}(y; \theta_1, 1) + (1 - w)\mathcal{N}(y; \theta_2, 1).$$

- $\boldsymbol{\phi} = (\theta_1, \theta_2, w)$.
- Infinitely many MLEs: $\boldsymbol{\phi}(\boldsymbol{\theta}) = (\theta, \theta, w), \boldsymbol{\phi}(\boldsymbol{\theta}) = (\theta_1, \theta, 0), \boldsymbol{\phi}(\boldsymbol{\theta}) = (\theta, \theta_2, 1)$.
- **In practice**: An auxiliary model $p_A$ containing an intractable $p$ is likely intractable.

# MCMC pdBIL [Gallant and McCulloch, 2009]

Straightforward use of the artificial likelihood

$$p_{A,n}(\mathbf{y}|\boldsymbol{\theta}) = \int p_A(\mathbf{y}|\boldsymbol{\phi}_n(\boldsymbol{\theta}, \mathbf{x}_{1:n})) \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta})\, \mathrm{d}\mathbf{x}_{1:n}.$$

Need a proposal $q(\boldsymbol{\theta}'|\boldsymbol{\theta})$. Important to use large $n$.

**Algorithm 2** MCMC pdBIL algorithm (see also Gallant and McCulloch (2009)).

Set $\boldsymbol{\theta}^0$
Simulate $\mathbf{x}^*_{1:n} \sim p(\cdot|\boldsymbol{\theta}^0)$
Compute $\boldsymbol{\phi}^0 = \arg\max_{\boldsymbol{\phi}\in\Phi} p_A(\mathbf{x}^*_{1:n}|\boldsymbol{\phi})$
**for** $i = 1$ **to** $T$ **do**
  Draw $\boldsymbol{\theta}^* \sim q(\cdot|\boldsymbol{\theta}^{i-1})$
  Simulate $\mathbf{x}^*_{1:n} \sim p(\cdot|\boldsymbol{\theta}^*)$
  Compute $\boldsymbol{\phi}(\mathbf{x}^*_{1:n}) = \arg\max_{\boldsymbol{\phi}} p_A(\mathbf{x}^*_{1:n}|\boldsymbol{\phi})$
  Compute $r = \min\left(1, \frac{p_A(\mathbf{y}|\boldsymbol{\phi}(\mathbf{x}^*_{1:n}))p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{i-1}|\boldsymbol{\theta}^*)}{p_A(\mathbf{y}|\boldsymbol{\phi}^{i-1})p(\boldsymbol{\theta}^{i-1})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{i-1})}\right)$
  **if** $\mathrm{uniform}(0,1) < r$ **then**
    $\boldsymbol{\theta}^i = \boldsymbol{\theta}^*$
    $\boldsymbol{\phi}^i = \boldsymbol{\phi}(\mathbf{x}^*_{1:n})$
  **else**
    $\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1}$
    $\boldsymbol{\phi}^i = \boldsymbol{\phi}^{i-1}$
  **end if**
**end for**

# Synthetic Likelihood [Wood, 2010]

- Difference to pdBIL = model the distribution of summary statistics
- Need $s(\mathbf{x})$.
- Auxiliary model:

$$p_A(s(\mathbf{y})|\boldsymbol{\phi}(\boldsymbol{\theta})) = \mathcal{N}(s(\mathbf{y}); \boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})),$$

  where $\phi(\boldsymbol{\theta}) = \{\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})\}$.

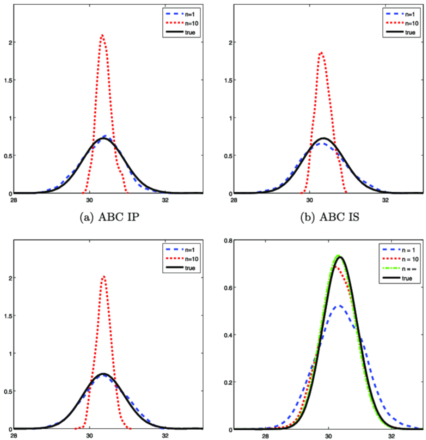
1. Draw $\boldsymbol{\theta}$.
2. Draw $\mathbf{x}_{1:n}|\boldsymbol{\theta}$. Compute $\{s(\mathbf{x}_1), \ldots, s(\mathbf{x}_n)\}$.
3. Find MLE $\hat{\phi}(\boldsymbol{\theta}) = \hat{\boldsymbol{\mu}}(\boldsymbol{\theta}), \hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$.
4. Approximate posterior ($n$ dependent):

$$p(\boldsymbol{\theta}|s(\mathbf{y})) \propto p(\boldsymbol{\theta}) p_A(s(\mathbf{y})|\hat{\boldsymbol{\phi}}(\boldsymbol{\theta})).$$

# Toy Example

- $\mathbf{y} = (y_1, \ldots, y_N)$. $N = 100$.
- $\mathbf{y} \overset{i.i.d.}{\sim} \text{Poisson}(\lambda = 30) = p(y|\lambda)$.
- Prior: $p(\lambda) = \text{Gamma}(\alpha = 30, \beta = 1)$.
- True posterior: $p(\lambda|\mathbf{y}) = \text{Gamma}(\alpha + \sum_{i=1}^{N} y_i, \beta + N)$.
- Auxiliary model: $p_A(y|\mu, \tau) = \mathcal{N}(y; \mu, \tau)$.
- For large $\lambda$, normal approximation is reasonable.

# Results

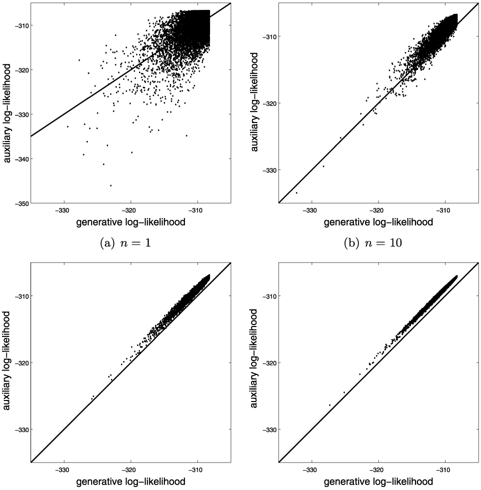


(a) ABC IP (b) ABC IS (c) ABC IL (d) pdBIL

- ABC IP: $\rho(s(\mathbf{x}), s(\mathbf{y})) = \sqrt{(\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{y}))^\top \mathbf{J}(\boldsymbol{\phi}(\mathbf{y}))^{-1}(\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{y}))}$
- ABC IL: $\rho(s(\mathbf{x}), s(\mathbf{y})) = \log p_A(\mathbf{y}|\boldsymbol{\phi}(\mathbf{y})) - \log p_A(\mathbf{y}|\boldsymbol{\phi}(\mathbf{x}))$
- ABC IS: $\rho(s(\mathbf{x}), s(\mathbf{y})) = \sqrt{\mathbf{S}_A(\mathbf{x}, \boldsymbol{\phi}(\mathbf{y}))^\top \mathbf{J}(\boldsymbol{\phi}(\mathbf{y}))^{-1} \mathbf{S}_A(\mathbf{x}, \boldsymbol{\phi}(\mathbf{y}))}$ where $\mathbf{S}_A(\mathbf{y}, \boldsymbol{\phi}) = \left(\frac{\partial \log p_A(\mathbf{y}|\boldsymbol{\phi})}{\partial \phi_1}, \dots, \frac{\partial \log p_A(\mathbf{y}|\boldsymbol{\phi})}{\partial \phi_{\dim(\boldsymbol{\phi})}}\right)$
- pdBIL: $p_{A,n}(\mathbf{y}|\boldsymbol{\theta}) = \int p_A(\mathbf{y}|\boldsymbol{\phi}_n(\boldsymbol{\theta}, \mathbf{x}_{1:n})) \prod_{i=1}^n p(\mathbf{x}_i|\boldsymbol{\theta}) \, \mathrm{d}\mathbf{x}_{1:n}$.

- High $n$ (replicates) helps in pdBIL and hurts others.

# High $n$ Is Good for pdBIL




(a) $n = 1$ (b) $n = 10$ (c) $n = 100$ (d) $n = 1000$

$$p_{A,n}(\mathbf{y}|\boldsymbol{\theta}) = \int p_A(\mathbf{y}|\boldsymbol{\phi}_n(\boldsymbol{\theta}, \mathbf{x}_{1:n})) \prod_{i=1}^n p(\mathbf{x}_i|\boldsymbol{\theta})\, d\mathbf{x}_{1:n}$$

- For high $n$, MLE map $\boldsymbol{\theta} \mapsto \boldsymbol{\phi}_n(\boldsymbol{\theta}, \mathbf{x}_{1:n})$ is less stochastic.
- $p_A$ (normal) is a good approx. to $p$ (Poisson) for high $\lambda$.
- Statistic $\overline{\mathbf{y}}$ is sufficient.

# Nonparametric Auxiliary Model

- Kernel density estimate (KDE) of $\mathbf{x}_{1:n}$. $\boldsymbol{\phi}(\boldsymbol{\theta}, \mathbf{x}_{1:n}) = \mathbf{x}_{1:n}$.
- Auxiliary likelihood:

$$p_A(\mathbf{y}|\boldsymbol{\phi}(\boldsymbol{\theta}, \mathbf{x}_{1:n})) = p_A(\mathbf{y}|\mathbf{x}_{1:n}) = \int K_\epsilon(\rho(\mathbf{y}, \mathbf{x}))p(\mathbf{x}|\boldsymbol{\theta})\,\mathrm{d}\mathbf{x}$$
$$\approx \frac{1}{n}\sum_{i=1}^{n} K_\epsilon(\rho(\mathbf{y}, \mathbf{x}_i)),$$

  where $K_\epsilon(\rho(\mathbf{y}, \mathbf{x}_i))$ is a normalized smoothing kernel with bandwidth $\epsilon$.
- Density estimation is generally difficult.
- Alternatively, KDE for $p_A(s(\mathbf{y})|\boldsymbol{\phi}(\boldsymbol{\theta}, \mathbf{x}_{1:n}))$ [Creel and Kristensen, 2013].

# Bonus 1: K2-ABC [Park et al., 2015]

- Use kernel MMD to define discrepancy $\rho$.
- ABC likelihood:

$$p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) = \int K_\epsilon(\kappa(\mathbf{x}, \mathbf{y}))p(\mathbf{x}|\boldsymbol{\theta})\,\mathrm{d}\mathbf{x},$$

  where $K_\epsilon(t) = \exp(-t^2/\epsilon)$ and $\kappa(\mathbf{x}, \mathbf{y}) = \widehat{\mathrm{MMD}}^2(\mathbf{x}, \mathbf{y})$.
- Positive definite kernel $k$.
- $\widehat{\mathrm{MMD}}^2(\mathbf{x}, \mathbf{y}) = \frac{1}{n(n-1)}\sum_{x \neq x' \in \mathbf{x}} k(x, x') + \frac{1}{n(n-1)}\sum_{y \neq y' \in \mathbf{y}} k(y, y') - \frac{2}{n^2}\sum_{x \in \mathbf{x}}\sum_{y \in \mathbf{y}} k(x, y)$
  - Map $\mathbf{x}, \mathbf{y}$ to an infinite-dim. space and compute the distance.
- If $k$ is characteristic, corresponds to using infinite-dim. summary statistics.

# Bonus 2: Full DR-ABC (Distribution Regression) [Mitrovic et al., 2016]

- Similar to ABC IP. Use predicted parameters to define discrepancy.
- **Training**: learn $f : \mathbf{x} \mapsto \boldsymbol{\theta}$ with kernel distribution regression on $\{(\mathbf{x}_l, \boldsymbol{\theta}_l)\}_{l=1}^{L} \sim p(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})$.
- Need a kernel on $\{\mathbf{x}_l\}_{l=1}^{L}$. Use $\mathsf{k}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\widehat{\mathrm{MMD}}^2(\mathbf{x},\mathbf{x}')}{2\sigma^2}\right)$.
- **ABC** likelihood:

$$p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) = \int K_\epsilon(\|f(\mathbf{x}) - f(\mathbf{y})\|_2^2)p(\mathbf{x}|\boldsymbol{\theta})\,\mathrm{d}\mathbf{x},$$

  where $K_\epsilon(t) = \exp(-t^2/\epsilon)$.
- $f(\mathbf{x})$ = summary statistics. Optimal under squared loss $\|\theta - \theta^*\|^2$.
- Use random Fourier features to approximate $\mathsf{k}(\mathbf{x}, \mathbf{x}')$.

# Cond DR-ABC [Mitrovic et al., 2016]

- Let $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$. Assume $p(\mathbf{x}|\boldsymbol{\theta}) = p(\mathbf{x}^{(1)}|\mathbf{x}^{(2)}, \boldsymbol{\theta})p(\mathbf{x}^{(2)}|\boldsymbol{\theta})$.
- **Claim**: better represent drawn $\mathbf{x}$ with $C_{\mathbf{x}^{(1)}|\mathbf{x}^{(2)}}$, a conditional mean embedding operator from $\mathbf{x}^{(2)}$ to $\mathbf{x}^{(1)}$.
- **Training**:
  - Generate $\{(\mathbf{x}_l, \boldsymbol{\theta}_l)\}_{l=1}^{L} \sim p(\boldsymbol{\theta})p(\mathbf{x}|\boldsymbol{\theta})$.
  - Compute $\{(C_{\mathbf{x}^{(1)}|\mathbf{x}^{(2)},l}, \boldsymbol{\theta}_l)\}_{l=1}^{L}$.
  - Learn $g : C_{\mathbf{x}^{(1)}|\mathbf{x}^{(2)}} \mapsto \boldsymbol{\theta}$.
- Need a kernel on $\{(C_{\mathbf{x}^{(1)}|\mathbf{x}^{(2)},l})\}_{l=1}^{L}$. Use linear kernel $\mathsf{k}(C, C') = \mathrm{tr}(C^\top C')$.
- **ABC** likelihood:

$$p_\epsilon(\mathbf{y}|\boldsymbol{\theta}) = \int K_\epsilon(\|g(C_{\mathbf{x}^{(1)}|\mathbf{x}^{(2)}}) - g(C_{\mathbf{y}^{(1)}|\mathbf{y}^{(2)}})\|_2^2)p(\mathbf{x}|\boldsymbol{\theta})\,\mathrm{d}\mathbf{x}.$$

- **Concerns**:
  - Computing $C_{\mathbf{x}^{(1)}|\mathbf{x}^{(2)},l}$ for each $l$ requires at least $O(|\mathbf{x}|^2)$.
  - How to split $\mathbf{x} = (\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$ in general?
  - $C_{\mathbf{x}^{(1)}|\mathbf{x}^{(2)},l}$ requires parameter tuning.

# References I


Blum, M. G. B., Nunes, M. A., Prangle, D., and Sisson, S. A. (2013).
A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation.
*Statistical Science*, 28(2):189–208.
arXiv: 1202.3819.

Creel, M. and Kristensen, D. (2013).
Indirect Likelihood Inference (revised).
UFAE and IAE Working Paper.

Drovandi, C. C., Pettitt, A. N., and Faddy, M. J. (2011).
Approximate Bayesian computation using indirect inference.
*Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):317–337.

# References II


Gallant, A. R. and McCulloch, R. E. (2009).
On the determination of general scientific models with application to asset pricing.
*Journal of the American Statistical Association*.

Gleim, A. and Pigorsch, C. (2013).
Approximate bayesian computation with indirect summary statistics.
*Draft paper: http://ect-pigorsch. mee. uni-bonn. de/data/research/papers*.

Marjoram, P., Molitor, J., Plagnol, V., and TavarÃ©, S. (2003).
Markov chain Monte Carlo without likelihoods.
*Proceedings of the National Academy of Sciences*, 100(26):15324–15328.

Mitrovic, J., Sejdinovic, D., and Teh, Y. W. (2016).
DR-ABC: Approximate Bayesian Computation with Kernel-Based Distribution Regression.
*arXiv:1602.04805 [cs, stat]*.
arXiv: 1602.04805.

# References III


Park, M., Jitkrittum, W., and Sejdinovic, D. (2015).
K2-ABC: Approximate Bayesian Computation with Kernel Embeddings.
*arXiv:1502.02558 [cs, stat]*.
arXiv: 1502.02558.

Reeves, R. and Pettitt, A. (2005).
A theoretical framework for approximate bayesian computation.
In *20th International Workshop on Statistical Modelling*, pages 393–396.

Wood, S. N. (2010).
Statistical inference for noisy nonlinear ecological dynamic systems.
*Nature*, 466(7310):1102–1104.