Approximate Bayesian computation with surrogate posteriors

Julyan Arbel

Inria Grenoble Rhône-Alpes, France

julyan.arbel@inria.fr www.julyanarbel.com

Florence Forbes

Inria Grenoble Rhône-Alpes

Hien Nguyen

La Trobe Univ Melbourne

Tin Nguyen

Univ Caen Normandie

Approximate Bayesian computation with surrogate posteriors

Julyan Arbel

Inria Grenoble Rhône-Alpes, France

julyan.arbel@inria.fr www.julyanarbel.com

Florence Forbes

Inria Grenoble Rhône-Alpes

Hien Nguyen

La Trobe Univ Melbourne

Tin Nguyen

Univ Caen Normandie

Outline

- Approximate Bayesian computation (ABC)
- Semi-automatic ABC [Fearnhead & Prangle, 2012]
- Surrogate posteriors [GLLiM, Deleforge et al, 2015]

- GLLiM-ABC procedures [Forbes et al, 2021]
- Theoretical properties
- Illustration to sound source localisation

- Conclusion

Data generating model

Prior: $\pi(\boldsymbol{\theta})$

Likelihood: $f_{\boldsymbol{\theta}}(\mathbf{z})$, such that $\mathbf{z} = \{z_1, \ldots, z_d\}$ can be simulated from $f_{\boldsymbol{\theta}}$

Inverse problem goal

Estimation of $\boldsymbol{\theta}$ given a unique observation $\mathbf{y} = \{y_1, \ldots, y_d\}$

Posterior: $\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\mathbf{y})$

What if $f_{\boldsymbol{\theta}}$ is not tractable, not available, too costly?

Approximate Bayesian Computation (ABC)

Goal: get a $\boldsymbol{\theta}$ sample from $\pi(\cdot \mid \mathbf{y})$

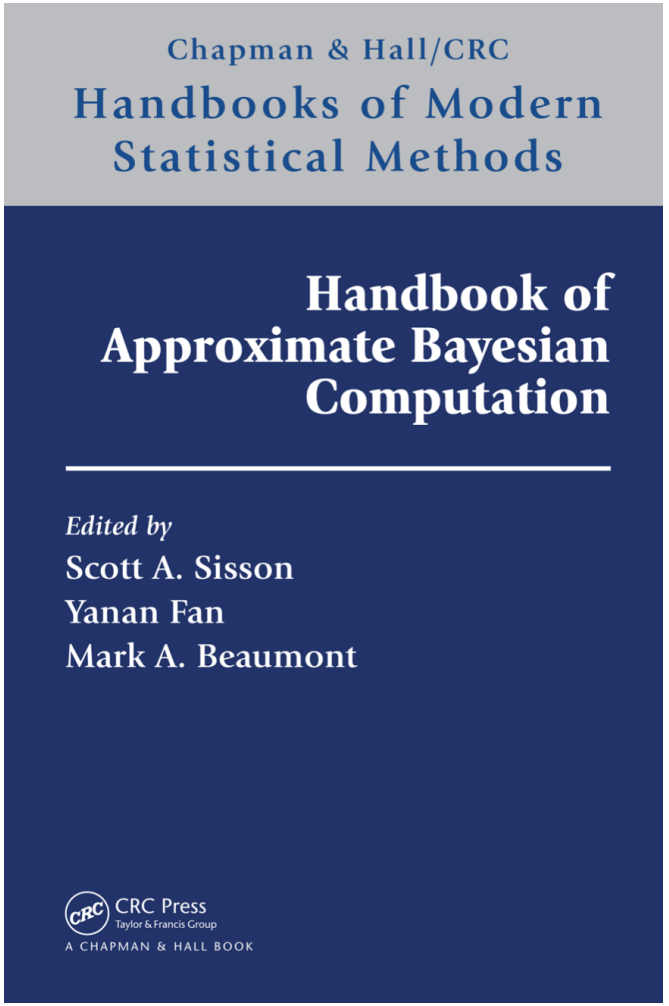
Simulate M iid $(\boldsymbol{\theta}_m\ ,\ \mathbf{z}_m)$ for $m = 1 : M$

1. $\boldsymbol{\theta}_m \sim \pi(\boldsymbol{\theta})$
2. $\mathbf{z}_m \mid \boldsymbol{\theta}_m \sim f_{\boldsymbol{\theta}_m}$
3. Keep $\boldsymbol{\theta}_m$ if $D(\mathbf{y}, \mathbf{z}_m) < \epsilon$

   $D(\mathbf{y}, \mathbf{z}_m) = \|\mathbf{y} - \mathbf{z}_m\|$ or $\|\mathbf{s}(\mathbf{y}) - \mathbf{s}(\mathbf{z}_m)\|$

   $\mathbf{s}$ is a summary statistic

$\longrightarrow$ Which choice for D? for $\mathbf{s}$?



Summary statistics and distances

1. Summary-based procedures

$\longrightarrow$ Replace $\|\mathbf{y} - \mathbf{z}_m\|$ by $\|\mathbf{s}(\mathbf{y}) - \mathbf{s}(\mathbf{z}_m)\|$

- Pros: Dimension reduction
- Cons: Arbitrary $\mathbf{s}$, loss of information

$\longrightarrow$ Semi-automatic ABC [Fearnhead & Prangle, 2012]: prelim learning step, d small

2. Data discrepancy-based procedures

$\longrightarrow$ Replace $\|\mathbf{y} - \mathbf{z}_m\|$ by distance between empirical distributions

- Maximum Mean Discrepancy [Park et al, 2016]
- Kullback–Leibler [Jiang et al, 2018]
- Classification accuracy [Gutmann et al, 2018]
- Wasserstein distance [Bernton & al 2019]
- Energy distance: [Nguyen & al 2020]

- Pros: Does not require summary statistics
- Cons: Requires moderately large samples, not available in inverse problems

Summary statistics and distances

1. Summary-based procedures

$\longrightarrow$ Replace $\|\mathbf{y} - \mathbf{z}_m\|$ by $\|\mathbf{s}(\mathbf{y}) - \mathbf{s}(\mathbf{z}_m)\|$

- Pros: Dimension reduction
- Cons: Arbitrary $\mathbf{s}$, loss of information

$\longrightarrow$ Semi-automatic ABC [Fearnhead & Prangle, 2012]: prelim learning step, d small

2. Data discrepancy-based procedures

$\longrightarrow$ Replace $\|\mathbf{y} - \mathbf{z}_m\|$ by distance between empirical distributions

- Maximum Mean Discrepancy [Park et al, 2016]
- Kullback–Leibler [Jiang et al, 2018]
- Classification accuracy [Gutmann et al, 2018]
- Wasserstein distance [Bernton & al 2019]
- Energy distance: [Nguyen & al 2020]

- Pros: Does not require summary statistics
- Cons: Requires moderately large samples, not available in inverse problems

Key idea

- **ABC posterior (Bayes' theorem):**

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbb{I}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, f_{\boldsymbol{\theta}}(\mathbf{z}) \, \mathrm{d}\mathbf{z} \quad \propto \int_{\mathcal{Y}} \mathbb{I}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, \pi(\boldsymbol{\theta} \mid \mathbf{z}) \, \pi(\mathbf{z}) \, \mathrm{d}\mathbf{z}$$

- **Our ABC posterior:** replace $D(\mathbf{y}, \mathbf{z})$ by $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$, where D is now a distance on distributions

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \int_{\mathcal{Y}} \mathbb{I}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\boldsymbol{\theta} \mid \mathbf{z}) \pi(\mathbf{z}) \mathrm{d}\mathbf{z}$$

Theorem 1 [Forbes et al] $q_\epsilon(\cdot \mid \mathbf{y}) \to \pi(\cdot \mid \mathbf{y})$ in total variation when $\epsilon \to 0$

Intuition of the proof:

When $\epsilon \to 0$, $\{\mathbf{z} \in \mathcal{Y}, D\left(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z})\right) \leq \epsilon\} \to \{\mathbf{z} \in \mathcal{Y}, \pi(\cdot \mid \mathbf{z}) = \pi(\cdot \mid \mathbf{y})\}$

$$\int_{\mathcal{Y}} \mathbb{I}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \pi\left(\mathbf{z}\right) \mathrm{d}\mathbf{z} \to \int_{\mathcal{Y}} \mathbb{I}_{\{\pi(\cdot|\mathbf{z}) = \pi(\cdot|\mathbf{y})\}} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \pi\left(\mathbf{z}\right) \mathrm{d}\mathbf{z} \propto \pi\left(\boldsymbol{\theta} \mid \mathbf{y}\right)$$

In practice: replace the unknown $\pi(\cdot \mid \mathbf{y})$ by an estimate

Key idea

- **ABC posterior (Bayes' theorem):**

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbb{I}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, f_{\boldsymbol{\theta}}(\mathbf{z}) \, \mathrm{d}\mathbf{z} \quad \propto \int_{\mathcal{Y}} \mathbb{I}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, \pi(\boldsymbol{\theta} \mid \mathbf{z}) \, \pi(\mathbf{z}) \, \mathrm{d}\mathbf{z}$$

- **Our ABC posterior:** replace $D(\mathbf{y}, \mathbf{z})$ by $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$, where D is now a distance on distributions

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \int_{\mathcal{Y}} \mathbb{I}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\boldsymbol{\theta} \mid \mathbf{z}) \pi(\mathbf{z}) \mathrm{d}\mathbf{z}$$

Theorem 1 [Forbes et al] $q_\epsilon(\cdot \mid \mathbf{y}) \to \pi(\cdot \mid \mathbf{y})$ in total variation when $\epsilon \to 0$

Intuition of the proof:

When $\epsilon \to 0$, $\{\mathbf{z} \in \mathcal{Y}, D\left(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z})\right) \leq \epsilon\} \to \{\mathbf{z} \in \mathcal{Y}, \pi(\cdot \mid \mathbf{z}) = \pi(\cdot \mid \mathbf{y})\}$

$$\int_{\mathcal{Y}} \mathbb{I}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \pi\left(\mathbf{z}\right) \mathrm{d}\mathbf{z} \to \int_{\mathcal{Y}} \mathbb{I}_{\{\pi(\cdot|\mathbf{z}) = \pi(\cdot|\mathbf{y})\}} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \pi\left(\mathbf{z}\right) \mathrm{d}\mathbf{z} \propto \pi\left(\boldsymbol{\theta} \mid \mathbf{y}\right)$$

In practice: replace the unknown $\pi(\cdot \mid \mathbf{y})$ by an estimate

Key idea

- **ABC posterior (Bayes' theorem):**

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbb{I}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, f_{\boldsymbol{\theta}}(\mathbf{z}) \, \mathrm{d}\mathbf{z} \quad \propto \int_{\mathcal{Y}} \mathbb{I}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, \pi(\boldsymbol{\theta} \mid \mathbf{z}) \, \pi(\mathbf{z}) \, \mathrm{d}\mathbf{z}$$

- **Our ABC posterior:** replace $D(\mathbf{y}, \mathbf{z})$ by $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$, where D is now a distance on distributions

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \int_{\mathcal{Y}} \mathbb{I}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\boldsymbol{\theta} \mid \mathbf{z}) \pi(\mathbf{z}) \mathrm{d}\mathbf{z}$$

Theorem 1 [Forbes et al] $q_\epsilon(\cdot \mid \mathbf{y}) \to \pi(\cdot \mid \mathbf{y})$ in total variation when $\epsilon \to 0$

Intuition of the proof:

When $\epsilon \to 0$, $\{\mathbf{z} \in \mathcal{Y}, D\left(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z})\right) \leq \epsilon\} \to \{\mathbf{z} \in \mathcal{Y}, \pi(\cdot \mid \mathbf{z}) = \pi(\cdot \mid \mathbf{y})\}$

$$\int_{\mathcal{Y}} \mathbb{I}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \pi\left(\mathbf{z}\right) \mathrm{d}\mathbf{z} \to \int_{\mathcal{Y}} \mathbb{I}_{\{\pi(\cdot|\mathbf{z}) = \pi(\cdot|\mathbf{y})\}} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \pi\left(\mathbf{z}\right) \mathrm{d}\mathbf{z} \propto \pi\left(\boldsymbol{\theta} \mid \mathbf{y}\right)$$

In practice: replace the unknown $\pi(\cdot \mid \mathbf{y})$ by an estimate

Key idea

- **ABC posterior (Bayes' theorem):**

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbb{I}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, f_{\boldsymbol{\theta}}(\mathbf{z}) \, \mathrm{d}\mathbf{z} \quad \propto \int_{\mathcal{Y}} \mathbb{I}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, \pi(\boldsymbol{\theta} \mid \mathbf{z}) \, \pi(\mathbf{z}) \, \mathrm{d}\mathbf{z}$$

- **Our ABC posterior:** replace $D(\mathbf{y}, \mathbf{z})$ by $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$, where D is now a distance on distributions

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \int_{\mathcal{Y}} \mathbb{I}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\boldsymbol{\theta} \mid \mathbf{z}) \pi(\mathbf{z}) \mathrm{d}\mathbf{z}$$

Theorem 1 [Forbes et al] $q_\epsilon(\cdot \mid \mathbf{y}) \to \pi(\cdot \mid \mathbf{y})$ in total variation when $\epsilon \to 0$

Intuition of the proof:

When $\epsilon \to 0$, $\{\mathbf{z} \in \mathcal{Y}, D\left(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z})\right) \leq \epsilon\} \to \{\mathbf{z} \in \mathcal{Y}, \pi(\cdot \mid \mathbf{z}) = \pi(\cdot \mid \mathbf{y})\}$

$$\int_{\mathcal{Y}} \mathbb{I}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \pi\left(\mathbf{z}\right) \mathrm{d}\mathbf{z} \to \int_{\mathcal{Y}} \mathbb{I}_{\{\pi(\cdot|\mathbf{z}) = \pi(\cdot|\mathbf{y})\}} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \pi\left(\mathbf{z}\right) \mathrm{d}\mathbf{z} \propto \pi\left(\boldsymbol{\theta} \mid \mathbf{y}\right)$$

In practice: replace the unknown $\pi(\cdot \mid \mathbf{y})$ by an estimate

Key idea

- **ABC posterior (Bayes' theorem):**

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbb{I}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, f_{\boldsymbol{\theta}}(\mathbf{z}) \, \mathrm{d}\mathbf{z} \quad \propto \int_{\mathcal{Y}} \mathbb{I}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, \pi(\boldsymbol{\theta} \mid \mathbf{z}) \, \pi(\mathbf{z}) \, \mathrm{d}\mathbf{z}$$

- **Our ABC posterior:** replace $D(\mathbf{y}, \mathbf{z})$ by $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$, where D is now a distance on distributions

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \int_{\mathcal{Y}} \mathbb{I}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\boldsymbol{\theta} \mid \mathbf{z}) \pi(\mathbf{z}) \mathrm{d}\mathbf{z}$$

Theorem 1 [Forbes et al] $q_\epsilon(\cdot \mid \mathbf{y}) \to \pi(\cdot \mid \mathbf{y})$ in total variation when $\epsilon \to 0$

Intuition of the proof:

When $\epsilon \to 0$, $\{\mathbf{z} \in \mathcal{Y}, D\left(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z})\right) \leq \epsilon\} \to \{\mathbf{z} \in \mathcal{Y}, \pi(\cdot \mid \mathbf{z}) = \pi(\cdot \mid \mathbf{y})\}$

$$\int_{\mathcal{Y}} \mathbb{I}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \pi\left(\mathbf{z}\right) \mathrm{d}\mathbf{z} \to \int_{\mathcal{Y}} \mathbb{I}_{\{\pi(\cdot|\mathbf{z}) = \pi(\cdot|\mathbf{y})\}} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \pi\left(\mathbf{z}\right) \mathrm{d}\mathbf{z} \propto \pi\left(\boldsymbol{\theta} \mid \mathbf{y}\right)$$

In practice: replace the unknown $\pi(\cdot \mid \mathbf{y})$ by an estimate

Surrogate posteriors as Gaussian mixtures

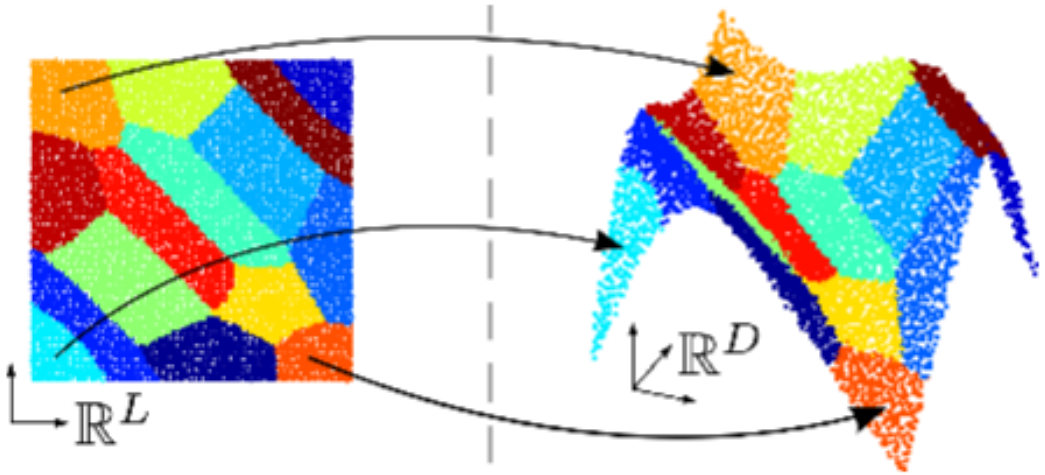
Gaussian Locally Linear mapping (GLLiM) model [Deleforge et al, 2015]

- provides a posterior for each $\mathbf{y}$ within parametric family $\{p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}), \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$
- captures link between $\mathbf{y}$ and $\boldsymbol{\theta}$ with mixture of K affine components

$$p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}) = \sum_{k=1}^{K} \eta_k(\mathbf{y}) \mathcal{N}(\boldsymbol{\theta}; \mathbf{A}_k \mathbf{y} + \mathbf{b}_k, \boldsymbol{\Sigma}_k)$$

Fit GLLiM model with a simulated learning set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n = 1 : N\}$

Parameters $\boldsymbol{\phi}^*_{K,N} = \{\pi^*_k, \mathbf{c}^*_k, \boldsymbol{\Gamma}^*_k, \mathbf{A}^*_k, \mathbf{b}^*_k, \boldsymbol{\Sigma}^*_k\}_{k=1}^K$ learned with EM algorithm



$\mathbf{y}$ space $\boldsymbol{\theta}$ space

Extended semi-automatic ABC: 3 variants of GLLiM-ABC

GLLiM surrogate posterior $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}) = \sum_{k=1}^{K} \eta_k(\mathbf{y}) \mathcal{N}(\boldsymbol{\theta}; \mathbf{A}_k \mathbf{y} + \mathbf{b}_k, \boldsymbol{\Sigma}_k)$

- **Variant 1**: approximate $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{z}]$ with $\mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]$

  $\rightarrow \sum_{k=1}^{K} \eta_k^*(\mathbf{y})(\mathbf{A}_k^* \mathbf{y} + \mathbf{b}_k^*)$

- **Variant 2**: add the log posterior variances $\mathrm{Var}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]$

  $\rightarrow \sum_{k=1}^{K} \eta_k^*(\mathbf{y}) \left[\boldsymbol{\Sigma}_k^* + (\mathbf{A}_k^* \mathbf{y} + \mathbf{b}_k^*)(\mathbf{A}_k^* \mathbf{y} + \mathbf{b}_k^*)^\top\right] - \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}] \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]^\top$

- **Variant 3**: use full $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N})$

  $\rightarrow$ requires a metric for Gaussian mixtures

  Mixture Wasserstein distance (MW2) [Delon & Desolneux 2020]

  L_2 distance

GLLiM-ABC procedures

1: **Inverse operator learning**
Apply GLLiM on simulated set $\mathcal{D}_N = \{(\boldsymbol{\theta}_i, \mathbf{z}_i), i = 1 : N\}$ to get $p_G(\boldsymbol{\theta} \mid \mathbf{z}, \boldsymbol{\phi}^*_{K,N})$

2: **Distances computation**
For another simulated set $\mathcal{E}_M = \{(\boldsymbol{\theta}_m, \mathbf{z}_m), m = 1 : M\}$ and a given observed $\mathbf{y}$, do

Vector summary statistics
Variant 1: *GLLiM-E-ABC: Compute GLLiM expectation*
Variant 2: *GLLiM-EV-ABC: Compute GLLiM expectation and log variance*
Compute standard distances between summary statistics

Functional summary statistics
Variant 3: *GLLiM-MW2-ABC: Compute* $MW_2(p_G(\cdot \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}), p_G(\cdot \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}))$
Variant 3': *GLLiM-L2-ABC: Compute* $L_2(p_G(\cdot \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}), p_G(\cdot \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}))$

3: **Sample selection**
Select the $\boldsymbol{\theta}_m$ values that correspond to distances under ϵ threshold (rejection ABC) or apply other ABC procedure (IS-ABC, MCMC-ABC, SMC-ABC)

Convergence of our ABC posterior when $\epsilon \to 0$, $K, N \to \infty$

ABC posterior $q_\epsilon^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbb{1}_{\left\{D\left(p^{K,N}(\cdot|\mathbf{y}), p^{K,N}(\cdot|\mathbf{z})\right) \leq \epsilon\right\}} f_{\boldsymbol{\theta}}(\mathbf{z})\, \mathrm{d}\mathbf{z}$

computed with surrogates $\{p^{K,N}(\cdot \mid \mathbf{y}) = p^K\left(\cdot \mid \mathbf{y}; \boldsymbol{\phi}_{K,N}^*\right) : \mathbf{y} \in \mathcal{Y}, K \in \mathbb{N}, N \in \mathbb{N}\}$
where:

- $\boldsymbol{\phi}_{K,N}^*$ MLE from simulated data $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n = 1 : N\}$ generated from joint
- p^K are K-component mixtures

Let D_H denote Hellinger distance. Then under compactness assumptions on true data generating process & additional "standard" assumptions:

Theorem 2 [Forbes et al, 2021] $D_H\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \xrightarrow[\epsilon \to 0, K, N \to \infty]{} 0$

Remark

- GLLiM involves multivariate unconstrained Gaussian distributions, does not satisfy the conditions: $p^{K,N}$ cannot be replaced by $p_G^{K,N}$
- Truncated Gaussian distributions with constrained parameters can meet the restrictions

Convergence of our ABC posterior when $\epsilon \to 0$, $K, N \to \infty$

ABC posterior $q_\epsilon^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbb{1}_{\left\{D\left(p^{K,N}(\cdot|\mathbf{y}), p^{K,N}(\cdot|\mathbf{z})\right) \leq \epsilon\right\}} f_{\boldsymbol{\theta}}(\mathbf{z})\, \mathrm{d}\mathbf{z}$

computed with surrogates $\{p^{K,N}(\cdot \mid \mathbf{y}) = p^K\left(\cdot \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}\right) : \mathbf{y} \in \mathcal{Y}, K \in \mathbb{N}, N \in \mathbb{N}\}$
where:

- $\boldsymbol{\phi}^*_{K,N}$ MLE from simulated data $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n = 1 : N\}$ generated from joint
- p^K are K-component mixtures

Let D_H denote Hellinger distance. Then under compactness assumptions on true data generating process & additional "standard" assumptions:

Theorem 2 [Forbes et al, 2021] $D_H\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \xrightarrow[\epsilon \to 0, K, N \to \infty]{} 0$

Remark

- GLLiM involves multivariate unconstrained Gaussian distributions, does not satisfy the conditions: $p^{K,N}$ cannot be replaced by $p_G^{K,N}$
- Truncated Gaussian distributions with constrained parameters can meet the restrictions

Convergence of our ABC posterior when $\epsilon \to 0$, $K, N \to \infty$

ABC posterior $q_\epsilon^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbb{1}_{\left\{D\left(p^{K,N}(\cdot|\mathbf{y}), p^{K,N}(\cdot|\mathbf{z})\right) \leq \epsilon\right\}} f_{\boldsymbol{\theta}}(\mathbf{z})\, \mathrm{d}\mathbf{z}$

computed with surrogates $\{p^{K,N}(\cdot \mid \mathbf{y}) = p^K\left(\cdot \mid \mathbf{y}; \boldsymbol{\phi}_{K,N}^*\right) : \mathbf{y} \in \mathcal{Y}, K \in \mathbb{N}, N \in \mathbb{N}\}$
where:

- $\boldsymbol{\phi}_{K,N}^*$ MLE from simulated data $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n = 1 : N\}$ generated from joint
- p^K are K-component mixtures

Let D_H denote Hellinger distance. Then under compactness assumptions on true data generating process & additional "standard" assumptions:

Theorem 2 [Forbes et al, 2021] $D_H\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \xrightarrow[\epsilon \to 0, K, N \to \infty]{} 0$

Remark

- GLLiM involves multivariate unconstrained Gaussian distributions, does not satisfy the conditions: $p^{K,N}$ cannot be replaced by $p_G^{K,N}$
- Truncated Gaussian distributions with constrained parameters can meet the restrictions

Illustration to sound source localisation

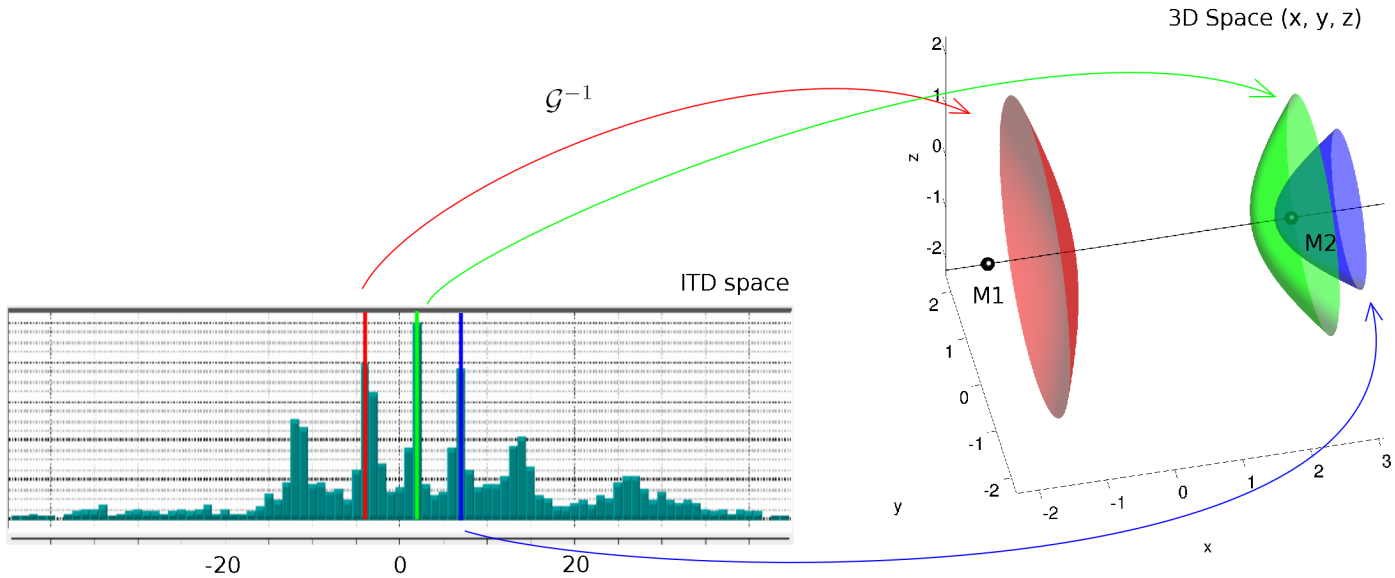
Multimodal posterior in inverse problem

Goal: find unknown location $\boldsymbol{\theta} = (x, y)$ of a sound source from two microphones at known positions $\mathbf{m}_1$ and $\mathbf{m}_2$ based on interaural time difference

$$\text{ITD}(\boldsymbol{\theta}) = \frac{1}{c}\left(\|\boldsymbol{\theta} - \mathbf{m}_1\|_2 - \|\boldsymbol{\theta} - \mathbf{m}_2\|_2\right)$$

Synthetic example in a 2D scene: $\mathbf{y} \sim \mathcal{S}_{10}(\text{ITD}(\boldsymbol{\theta})\mathbf{1}_d, \sigma^2\mathbf{I}_d, \nu)$ with $\sigma^2 = 0.01$ and $\nu = 3$

$\longrightarrow$ Posterior distribution that concentrates around two hyperboloids



Sound source localisation : two pairs of microphones & setting

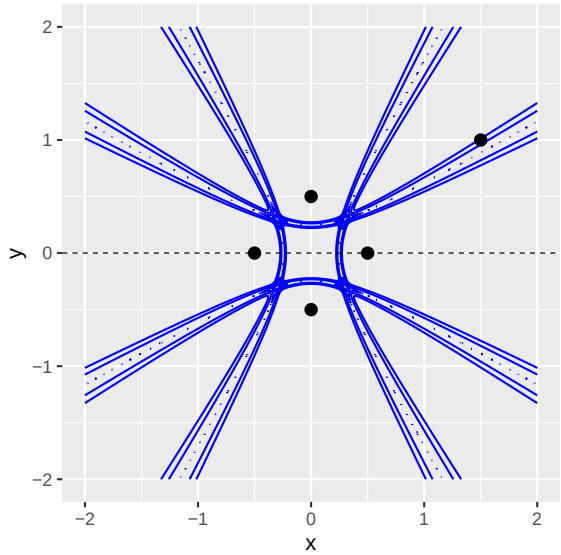
Two pairs of microphones:

True source position in $\boldsymbol{\theta} = (1.5, 1)$ either captured by the first or the second microphone pair

Likelihood is a mixture of 2 single-pair components

Posterior exhibits 4 symmetric hyperbolas

Contours of the true posterior:



Setting:

- GLLiM: learning set $N = 10^5$, K number of Gaussians set to 20, isotropic constraint (**xLLiM** package [Perthame et al, 2017])
- Rejection ABC: simulations $M = 10^6$, ϵ is 0.1% quantile of distance values $\rightarrow 10^3$ samples

Comparison of different (rejection ABC) procedures :

- Semi-automatic ABC (**abctools** R package [Nunes and Prangle, 2015])
- GLLiM-E-ABC: GLLiM expectations as summary stats (**abc** package [Csillery et al, 2012])
- GLLiM-EV-ABC: GLLiM expectations and log variances (**abc** package)
- GLLiM-L2-ABC and GLLiM-MW2-ABC (**transport** package [Schuhmacher et al, 2020])

Sound source localisation : two pairs of microphones & setting

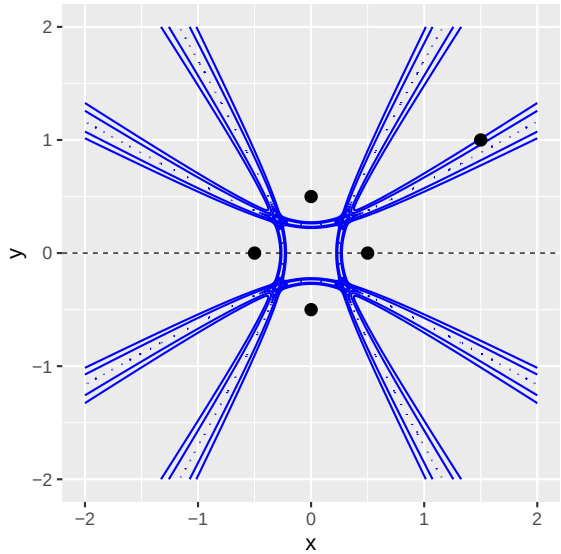
Two pairs of microphones:

True source position in $\boldsymbol{\theta} = (1.5, 1)$ either captured by the first or the second microphone pair

Likelihood is a mixture of 2 single-pair components

Posterior exhibits 4 symmetric hyperbolas

Contours of the true posterior:



Setting:

- GLLiM: learning set $N = 10^5$, K number of Gaussians set to 20, isotropic constraint (**xLLiM** package [Perthame et al, 2017])
- Rejection ABC: simulations $M = 10^6$, ϵ is 0.1% quantile of distance values $\rightarrow 10^3$ samples

Comparison of different (rejection ABC) procedures :

- Semi-automatic ABC (**abctools** R package [Nunes and Prangle, 2015])
- GLLiM-E-ABC: GLLiM expectations as summary stats (**abc** package [Csillery et al, 2012])
- GLLiM-EV-ABC: GLLiM expectations and log variances (**abc** package)
- GLLiM-L2-ABC and GLLiM-MW2-ABC (**transport** package [Schuhmacher et al, 2020])

Sound source localisation : two pairs of microphones & setting

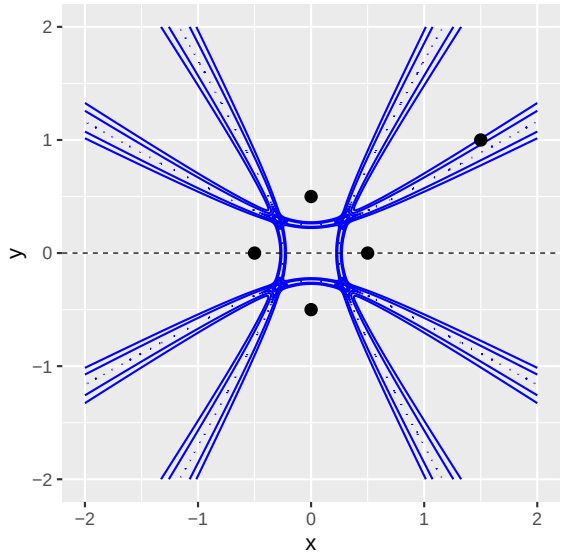
Two pairs of microphones:

True source position in $\boldsymbol{\theta} = (1.5, 1)$ either captured by the first or the second microphone pair

Likelihood is a mixture of 2 single-pair components

Posterior exhibits 4 symmetric hyperbolas

Contours of the true posterior:



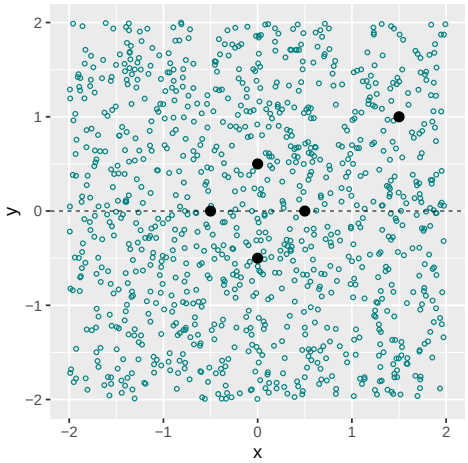
Setting:

- GLLiM: learning set $N = 10^5$, K number of Gaussians set to 20, isotropic constraint (**xLLiM** package [Perthame et al, 2017])
- Rejection ABC: simulations $M = 10^6$, ϵ is 0.1% quantile of distance values $\rightarrow 10^3$ samples

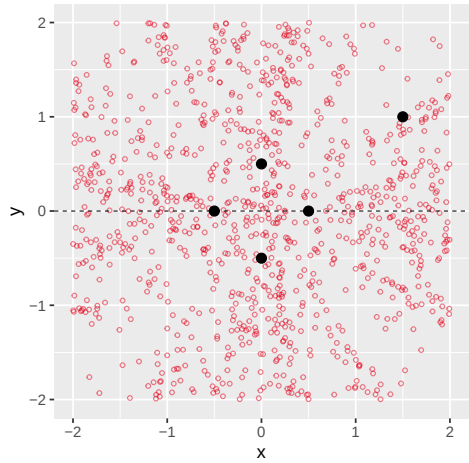
Comparison of different (rejection ABC) procedures :

- Semi-automatic ABC (**abctools** R package [Nunes and Prangle, 2015])
- GLLiM-E-ABC: GLLiM expectations as summary stats (**abc** package [Csillery et al, 2012])
- GLLiM-EV-ABC: GLLiM expectations and log variances (**abc** package)
- GLLiM-L2-ABC and GLLiM-MW2-ABC (**transport** package [Schuhmacher et al, 2020])

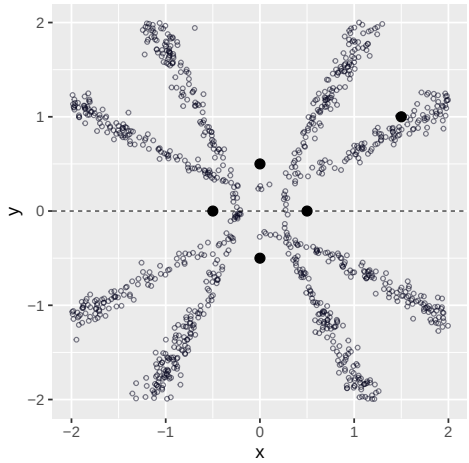
Sound source localisation : selected samples



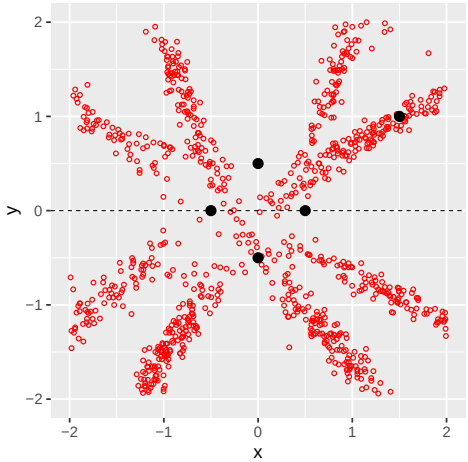
Semi-automatic ABC



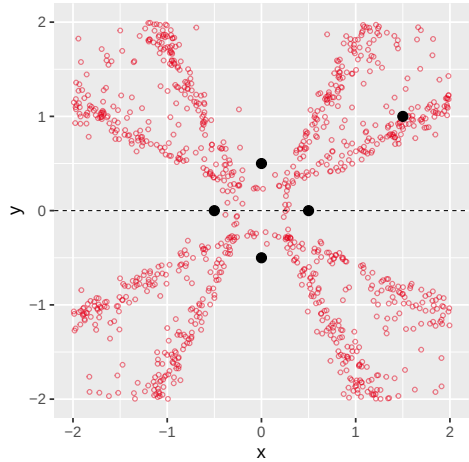
GLLiM-E-ABC



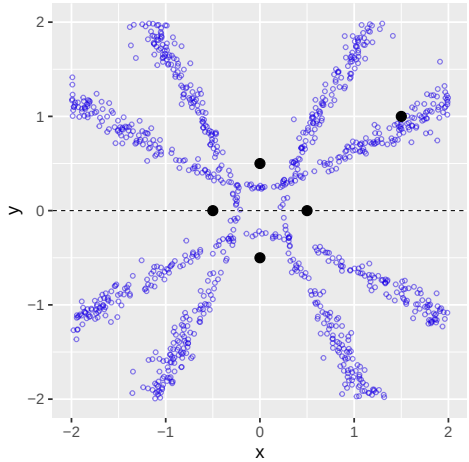
GLLiM-MW2-ABC



GLLiM mixture



GLLiM-EV-ABC



GLLiM-L2-ABC

Conclusion

Extended *semi-automatic ABC* with surrogate posteriors in place of summary statistics

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbb{I}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z}))\leq\epsilon\}} f_{\boldsymbol{\theta}}(\mathbf{z}) \mathrm{d}\mathbf{z}$$

Requirements:

1. Tractable, scalable model to learn surrogates (GLLiM up to $d = 100$-1000)
2. Metric between distributions: e.g. L_2, MW_2

First results and conclusions:

- No need to choose summary statistics
- (Restricted) convergence result to the true posterior
- Outperforms competitors for multimodal posteriors

Perspectives:

- Other learning schemes than GLLiM such as normalizing flows
- Extension to more-than-one-observation setting
- GLLiM-ABC$^{\mathbf{p}}$

Thank you for your attention!

Preprint: Forbes, Nguyen, Nguyen, Arbel. ABC with surrogate posteriors.

https://hal.archives-ouvertes.fr/hal-03139256

Bernton, Jacob, Gerber, Robert (2019). Inference in generative models using the Wasserstein distance. JRSS B.

Deleforge, Forbes, Horaud (2015). High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables. Statistics & Computing.

Delon, Desolneux (2020). A Wasserstein-type distance in the space of Gaussian Mixture Models. SIAM Journal on Imaging Sciences.

Fearnhead, Prangle (2012). Constructing summary statistics for ABC: semi-automatic ABC. JRSS B.

Jiang, Wu, Zheng, Wong (2017). Learning summary statistics for ABC via Deep Neural Network. Stat Sin.

Nguyen, Arbel, Lü, Forbes (2020). Approximate Bayesian Computation Via the Energy Statistic. IEEE Access.

Park, Jitkrittum, Sejdinovic (2016). K2-ABC: ABC with kernel embeddings. AISTATS.

Rubio, Johansen (2013). A simple approach to maximum intractable likelihood estimation. EJS.

Convergence of the ABC quasi-posterior: Rejection ABC

Goal: sample approximately from $\pi(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\mathbf{y})$ using $D(\mathbf{y}, \mathbf{z})$ $(D(\mathbf{s}(\mathbf{y}), \mathbf{s}(\mathbf{z})))$

Rejection ABC: replace intractable $f_{\boldsymbol{\theta}}$ by: $L_\epsilon(\mathbf{y}, \theta) = \int_{\mathcal{Y}} \mathbb{1}_{\{D(\mathbf{y},\mathbf{z})<\epsilon\}} f_{\boldsymbol{\theta}}(\mathbf{z})\, \mathrm{d}\mathbf{z}$

$\longrightarrow$ ABC quasi-posterior: $\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbb{1}_{\{D(\mathbf{y},\mathbf{z})<\epsilon\}} f_{\boldsymbol{\theta}}(\mathbf{z})\, d\mathbf{z}$

Convergence of the quasi-posterior to $\pi(\boldsymbol{\theta} \mid \mathbf{y})$: intuition of the proof

when $\epsilon \to 0$ then $D(\mathbf{y}, \mathbf{z}) \to 0$ so $\mathbf{z} \to \mathbf{y}$ and $\{\mathbf{z} \in \mathcal{Y},\ D(\mathbf{y}, \mathbf{z}) < \epsilon\} \to \{\mathbf{y}\}$

$$\pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbb{1}_{\{D(\mathbf{y},\mathbf{z})<\epsilon\}} f_{\boldsymbol{\theta}}(\mathbf{z})\, d\mathbf{z} \ \to\ \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbb{1}_{\{\mathbf{z}=\mathbf{y}\}} f_{\boldsymbol{\theta}}(\mathbf{z})\, d\mathbf{z} \ \to\ \pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\mathbf{y})$$

Details in [Rubio & Johansen 2013, Prangle et al, 2018, Bernton et al, 2019]

Semi-automatic ABC [Fearnhead & Prangle, 2012]

The posterior mean is the optimal (quadratic loss) summary : $\mathbf{s}(\mathbf{z}) = \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{z}]$

$\rightarrow$ Use a preliminary linear regression step to learn an approximation of $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{z}]$ as a function of $\mathbf{z}$

from $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n = 1 : N\}$ simulated from the true joint distribution

- **Variant 1**: replace linear regression by neural networks ... [Jiang et al, 2017, Wiqvist et al, 2019]

- **Variant 2**: add extra higher order moments (eg variances) in $\mathbf{s}$

  A natural idea mentioned (not implemented) in [Jiang et al, 2017]

$\rightarrow$ Requires a procedure able to provide posterior moments at low cost

- **Variant 3**: replace $\mathbf{s}(\mathbf{z})$ by an approximation (surrogate) of $\pi(\boldsymbol{\theta} \mid \mathbf{z})$

  Requires

  $\rightarrow$ a learning procedure able to provide tractable approximate posteriors at low cost: Gaussian Locally Linear Mapping [Deleforge et al, 2015]

  $\rightarrow$ a tractable metric between distributions to compare them