Approximate Bayesian computation with surrogate posteriors

Florence Forbes

Inria Grenoble Rhône-Alpes
France
Florence.Forbes@inria.fr

Joint work with

Hien Duy Nguyen (La Trobe University Melbourne),
Trung Tin Nguyen (Université Caen Normandie)
Julyan Arbel (Inria Grenoble Rhône-Alpes)

April 12, 2021

A data generating model

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

Likelihood: $f_{\boldsymbol{\theta}}(\mathbf{z})$

$\longrightarrow \mathbf{z} = \{z_1, \ldots, z_d\}$ can be simulated from $f_{\boldsymbol{\theta}}$

Goal: Estimation of $\boldsymbol{\theta}$ given some observed $\mathbf{y} = \{y_1, \ldots, y_d\}$

$$\text{Posterior: } \pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\mathbf{x})$$

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

Approximate Bayesian Computation (ABC)

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

Simulate M *i.i.d.* $(\boldsymbol{\theta}_m \, , \, \mathbf{z}_m)$ **for** $m = 1 \dots M$

$$\boldsymbol{\theta}_m \sim \pi(\boldsymbol{\theta})$$

$$\mathbf{z}_m \sim f_{\boldsymbol{\theta}_m}$$

If $D(\mathbf{y}, \mathbf{z}_m) < \epsilon$ then keep $\boldsymbol{\theta}_m$ **[Rejection ABC]**

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

$\mathbf{s}$ is a summary statistic

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

Difficult to select a summary statistic in general

Semi-automatic ABC [Fearnhead & Prangle 2012]

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

$\rightarrow$ Use a preliminary linear regression step to learn an approximation of $\mathbb{E}[\boldsymbol{\theta}|\mathbf{z}]$ 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 $\pi(\boldsymbol{\theta}|\mathbf{z})$

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

Surrogate posteriors as mixtures of Gaussians

The Gaussian Locally Linear mapping (GLLiM) model : an inverse regression approach that

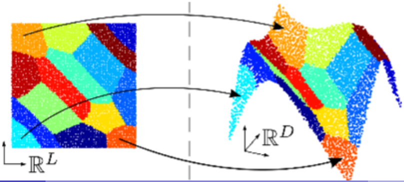
- aims at capturing the link between $\mathbf{y}$ and $\boldsymbol{\theta}$ with a mixture of K affine components
- provides for each $\mathbf{y}$ a posterior within a parametric family $\{p_G(\boldsymbol{\theta}|\mathbf{y};\boldsymbol{\phi}), \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$

$$\boldsymbol{\phi} = \{\pi_k, \mathbf{c}_k, \boldsymbol{\Gamma}_k, \mathbf{A}_k, \mathbf{b}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K \quad \text{and} \quad p_G(\boldsymbol{\theta}|\mathbf{y};\boldsymbol{\phi}) = \sum_{k=1}^K \eta_k(\mathbf{y})\,\mathcal{N}(\boldsymbol{\theta};\mathbf{A}_k\mathbf{y}+\mathbf{b}_k, \boldsymbol{\Sigma}_k)$$

mixture components: $\mathcal{N}(.;\boldsymbol{\mu},\boldsymbol{\Sigma})$ Gaussian pdf with mean $\boldsymbol{\mu}$, covariance $\boldsymbol{\Sigma}$

mixture weights: $\eta_k(\mathbf{y}) = \dfrac{\pi_k \mathcal{N}(\mathbf{y};\mathbf{c}_k,\boldsymbol{\Gamma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{y};\mathbf{c}_j,\boldsymbol{\Gamma}_j)}$

Fit a GLLiM model to a learning set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n = 1:N\}$ simulated from the true joint distribution: parameters $\boldsymbol{\phi}$ learned with an EM algorithm $\boldsymbol{\phi}^*_{K,N} = \{\pi^*_k, \mathbf{c}^*_k, \boldsymbol{\Gamma}^*_k, \mathbf{A}^*_k, \mathbf{b}^*_k, \boldsymbol{\Sigma}^*_k\}_{k=1}^K$



Extended semi-automatic ABC: GLLiM-ABC

GLLiM surrogate posteriors for each $\mathbf{y}$, $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N})$ with $\boldsymbol{\phi}^*_{K,N}$ independent of $\mathbf{y}$

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

- **Variant 1**: approximate $\mathbb{E}[\boldsymbol{\theta}|\mathbf{z}]$ with $\mathbb{E}_G[\boldsymbol{\theta}|\mathbf{y};\boldsymbol{\phi}^*_{K,N}] = \sum_{k=1}^{K} \eta^*_k(\mathbf{y})(\mathbf{A}^*_k\mathbf{y} + \mathbf{b}^*_k)$
- **Variant 2**: add the log posterior variances from

$$\mathrm{Var}_G[\boldsymbol{\theta}|\mathbf{y};\boldsymbol{\phi}^*_{K,N}] = \sum_{k=1}^{K} \eta^*_k(\mathbf{y}) \left[\boldsymbol{\Sigma}^*_k + (\mathbf{A}^*_k\mathbf{y} + \mathbf{b}^*_k)(\mathbf{A}^*_k\mathbf{y} + \mathbf{b}^*_k)^\top\right]$$
$$- (\sum_{k=1}^{K} \eta^*_k(\mathbf{y})(\mathbf{A}^*_k\mathbf{y} + \mathbf{b}^*_k))(\sum_{k=1}^{K} \eta^*_k(\mathbf{y})(\mathbf{A}^*_k\mathbf{y} + \mathbf{b}^*_k))^\top$$

- **Variant 3**: use full $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}) \rightarrow$ requires a metric for Gaussian mixtures

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

  $\rightarrow$ L_2 distance

GLLIM-ABC procedures

1: **Inverse operator learning.** Apply GLLiM on $\mathcal{D}_N$ to get for any $\mathbf{z}$ $p_G(\boldsymbol{\theta} \mid \mathbf{z}, \boldsymbol{\phi}^*_{K,N})$ as a first approximation of the true posterior $\pi(\boldsymbol{\theta} \mid \mathbf{z})$

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 one of the following for each m:

Vector summary statistics:
GLLiM-E-ABC: Compute summary $s_1(\mathbf{z}_m) = \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}]$
GLLiM-EV-ABC: Compute $s_1(\mathbf{z}_m)$ *and* $s_2(\mathbf{z}_m)$ *the GLLiM posterior log-variances*
Compute standard distances between summary statistics

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

3: **Sample selection.** Select the $\boldsymbol{\theta}_m$ values that correspond to distances under an ϵ threshold (rejection ABC) or apply some standard ABC procedure

4: **Sample use.** Use produced $\boldsymbol{\theta}$ values to get a closer approximation of $\pi(\boldsymbol{\theta}|\mathbf{y})$

Theoretical properties

- **A new quasi-posterior:** $q_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \propto \int_{\mathcal{Y}} \mathbb{1}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z}))\leq\epsilon\}} \pi(\boldsymbol{\theta}|\mathbf{z})\pi(\mathbf{z})d\mathbf{z}$

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

In practice: replace the unknown $\pi(\cdot|\mathbf{y})$ by a tractable approximation

- **ABC quasi-posterior with surrogate posteriors** $\{p^{K,N}(\cdot|\mathbf{y})\colon \mathbf{y}\in\mathcal{Y}, K\in\mathbb{N}, N\in\mathbb{N}\}$

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

Result [FF et al, Theorem 2] : $\epsilon \to 0$, $K, N \to \infty$

The Hellinger distance $D_H\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ converges to 0

- in some measure λ, with respect to $\mathbf{y} \in \mathcal{Y}$
- in probability, with respect to the sample $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n = 1 : N\}$

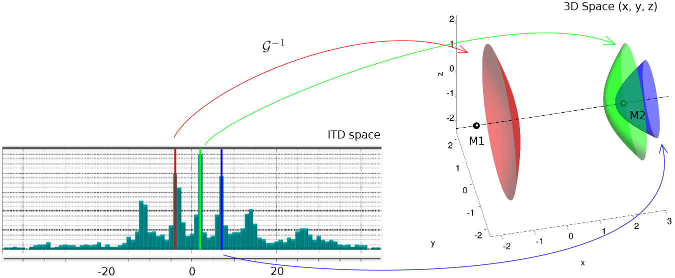
Restrictions:

- $p^{K,N}$ cannot be replaced by GLLiM $p_G^{K,N}$
- Truncated Gaussian distributions with constrained parameters can meet the restrictions

Illustration on sound source localisation

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

Sound localization cue: Interaural time difference $ITD(\theta) = \frac{1}{c}(||\theta - \mathbf{m}_1||_2 - ||\theta - \mathbf{m}_2||_2)$ but a whole hyperboloid of solutions



Synthetic example in a 2D scene: $\mathbf{y} \sim \mathcal{S}_{10}(F(\boldsymbol{\theta})\mathbf{1}_d, \sigma^2\mathbf{I}_d, \nu)$ with $F(\boldsymbol{\theta}) = |(||\boldsymbol{\theta} - \mathbf{m}_1||_2 - ||\boldsymbol{\theta} - \mathbf{m}_2||_2)|$

$\mathbf{y}$ is a $d = 10$-dimensional Student realization with $\sigma^2 = 0.01$ and $\nu = 3$

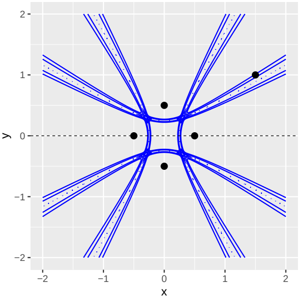
$\longrightarrow$ Posterior distribution that concentrates around two hyperboloids

Sound source localisation : two pairs of microphones

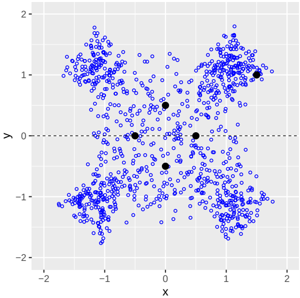
True source position in $\boldsymbol{\theta} = (1.5, 1)$ either captured by the first or the second microphone pair

Likelihood: an equal-weight mixture of 2 single-pair components

Posterior exhibits 4 symmetric hyperbolas



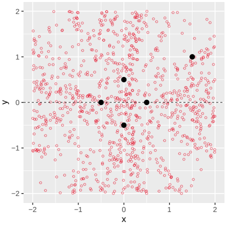
Contours of the true posterior



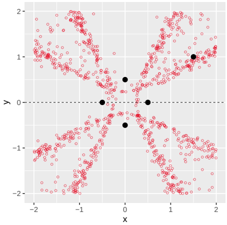
Metropolis–Hastings sample

Sound source localisation : selected samples

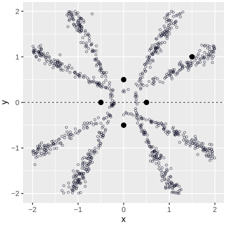
GLLiM $N = 10^5, K = 20$; Rejection ABC $M = 10^6$, $\epsilon = 0.1\%$ quantile (1000 values)



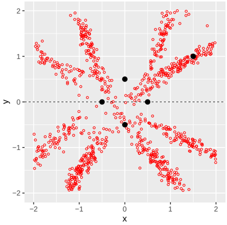
GLLiM-E-ABC



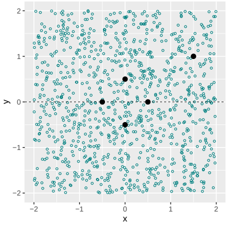
GLLiM-EV-ABC



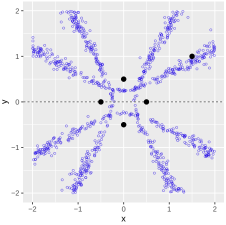
GLLiM-MW2-ABC



GLLiM mixture



semi-automatic ABC



GLLiM-L2-ABC

Conclusion

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

Requirements:

- A tractable, scalable model to learn the surrogates : e.g. GLLiM up to $d = 100, d = 1000$; can deal with missing data; latent variables
- A metric between distributions: e.g. L_2, MW_2

First results and conclusions:

- No need to choose summary statistics
- A (restricted) convergence result to the true posterior
- Satisfying performance when posteriors are multimodal
- Surrogate posterior quality seems not critical
- Wasserstein-based distance seems more robust than L_2

Perspectives

Short term improvements/ Future work:

- GLLiM use & implementation: information criterion to select K, test with higher d
- GLLiM-ABC: assess/compare computation costs
- More complete experiments and illustrations
- Other metrics between distributions
- Other learning scheme than GLLiM (Mixture density networks, Invertible NN)
- Other ABC scheme than rejection ABC (IS ABC, MCMC ABC, SMC ABC etc.)
- Refine choice of the threshold level
- Extension to i.i.d observations

Special thanks to: Guillaume Kon Kam King, Benoit Kugler and Sylvain Douté

Thank you for your attention !

Paper: F. Forbes, H. Nguyen, T. Nguyen, J. Arbel, ABC with surrogate posteriors

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

References

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

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

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

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

Gutmann, M., Dutta, R., Kaski, S., and Corander, J. (2018). Likelihood-free inference via classification. Statistics & Computing.

Jiang, B., et al. (2017). Learning summary statistics for ABC via Deep Neural Network. Statistica Sinica.

Jiang, B., et al. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy, AISTATS.

Nguyen, H. D., Arbel, J., Lü, H. & Forbes, F. (2020). Approximate Bayesian Computation Via the Energy Statistic.IEEE Access.

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

Prangle, D., Everitt, R. G. & Kypraios, T. (2018). A rare event approach to high-dimensional ABC. Statistics & Computing.

Rubio, F. & Johansen, A. M. (2013). A simple approach to maximum intractable likelihood estimation. Electronic Journal of Statistics.

Wiqvist, S., Mattei, P.-A., Picchini, U. & Frellsen, J. (2019). Partially exchangeable networks and architectures for learning summary statistics in ABC. ICML.

Appendix: GLLiM model hierarchical definition

$$\mathbf{y} = \sum_{k=1}^{K} \mathbb{1}_{(z=k)}(\mathbf{A}'_k \mathbf{x} + \mathbf{b}'_k + \mathbf{E}'_k)$$

$\mathbf{y} \in R^d$, $\mathbf{x}(\boldsymbol{\theta}) \in R^L$ with $d >> L$, $\mathbb{1}$ Indicator function, $\mathbf{A}'_k$ $d \times L$ matrix, $\mathbf{b}'_k$ d-dim vector

$\mathbf{E}'_k$: observation noise in $\mathbb{R}^d$ and reconstruction error, Gaussian, centered, independent on $\mathbf{x}$, $\mathbf{y}$, and z

$$p(\mathbf{y}|\mathbf{x}, z = k; \boldsymbol{\phi}') = \mathcal{N}(\mathbf{y}; \mathbf{A}'_k \mathbf{x} + \mathbf{b}'_k, \boldsymbol{\Sigma}'_k)$$

- Affine transformations are local: mixture of K Gaussians

$$\begin{aligned} p(\mathbf{x}|z = k; \boldsymbol{\phi}') &= \mathcal{N}(\mathbf{x}; \mathbf{c}'_k, \boldsymbol{\Gamma}'_k) \\ p(z = k; \boldsymbol{\phi}') &= \pi'_k \end{aligned}$$

- The set of all model parameters is:

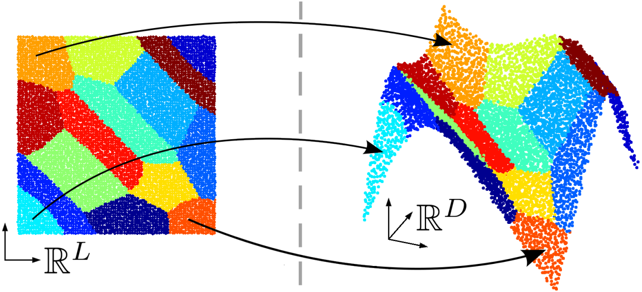
$$\boldsymbol{\phi}' = \{\mathbf{c}'_k, \boldsymbol{\Gamma}'_k, \pi'_k, \mathbf{A}'_k, \mathbf{b}'_k, \boldsymbol{\Sigma}'_k\}_{k=1}^{K}$$

Usually $\boldsymbol{\Sigma}'_k = \sigma^2 \mathbf{I}_d$ for $k = 1 \dots K$ (isotropic reconstruction error)

Appendix : GLLiM Geometric Interpretation

This model induces a partition of $\mathbb{R}^L$ into K regions $\mathcal{R}_k$ where the transformation τ_k is the most probable.

If $|\boldsymbol{\Gamma}'_1| = \dots = |\boldsymbol{\Gamma}'_K|$: $\{\mathcal{R}_k, k = 1 \dots K\}$ define a Voronoi diagram of centroids $\{\mathbf{c}'_k, k = 1 \dots K\}$ (Mahalanobis distance $||.||_{\boldsymbol{\Gamma}'}$).



$$p_G(\mathbf{y}|\mathbf{x}, \boldsymbol{\phi}') = \sum_{k=1}^{K} \eta'_k(\mathbf{x}) \mathcal{N}(\mathbf{y}; \mathbf{A}'_k \mathbf{x} + \mathbf{b}'_k, \boldsymbol{\Sigma}'_k) \quad \text{with } \eta'_k(\mathbf{x}) = \frac{\pi'_k \mathcal{N}(\mathbf{x}; \mathbf{c}'_k, \boldsymbol{\Gamma}'_k)}{\sum_{j=1}^{K} \pi'_j \mathcal{N}(\mathbf{x}; \mathbf{c}'_j, \boldsymbol{\Gamma}'_j)}$$

$$p_G(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}) = \sum_{k=1}^{K} \eta_k(\mathbf{y}) \mathcal{N}(\mathbf{x}; \mathbf{A}_k \mathbf{y} + \mathbf{b}_k, \boldsymbol{\Sigma}_k) \quad \text{with } \eta_k(\mathbf{y}) = \frac{\pi_k \mathcal{N}(\mathbf{y}; \mathbf{c}_k, \boldsymbol{\Gamma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{y}; \mathbf{c}_j, \boldsymbol{\Gamma}_j)}$$

Appendix : GLLiM link between ϕ and ϕ'

$$
\begin{aligned}
\mathbf{c}_k =& \mathbf{A}'_k \mathbf{c}'_k + \mathbf{b}'_k \\
\mathbf{\Gamma}_k =& \mathbf{\Sigma}'_k + \mathbf{A}'_k \mathbf{\Gamma}'_k \mathbf{A}'^{\top}_k \\
\mathbf{\Sigma}_k =& \left(\mathbf{\Gamma}'^{-1}_k + \mathbf{A}'^{\top}_k \mathbf{\Sigma}'^{-1}_k \mathbf{A}'_k \right)^{-1} \\
\mathbf{A}_k =& \mathbf{\Sigma}_k \mathbf{A}'^{\top}_k \mathbf{\Sigma}'^{-1}_k \\
\mathbf{b}_k =& \mathbf{\Sigma}_k \left(\mathbf{\Gamma}'^{-1}_k \mathbf{c}'_k - \mathbf{A}'^{\top}_k \mathbf{\Sigma}'^{-1}_k \mathbf{b}'_k \right)
\end{aligned}
$$

The number of parameters depends on the GLLiM variant but is in $\mathcal{O}(dKL)$

If diagonal covariances $\mathbf{\Sigma}_k$, the number of parameters is $K - 1 + K(L + L(L+1)/2 + dL + 2d)$

$\rightarrow$ for $K = 100$, $L = 4$ and $d = 10$ leads to 7499 parameters and to 61499 parameters if $d = 100$.

- **Optimal transport-based distance** [Delon & Desolneux 2020]

Quadratic cost Wasserstein distance between $g_1 = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $g_2 = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$:

$$\mathsf{W}_2^2(g_1, g_2) = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2 + \mathrm{trace}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2\left(\boldsymbol{\Sigma}_1^{1/2}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{1/2}\right)^{1/2}\right)$$

Mixture Wasserstein distance (MW2) between two Gaussian mixtures $f_1 = \sum_{k=1}^{K_1} \pi_{1k}\, g_{1k}$ and $f_2 = \sum_{k=1}^{K_2} \pi_{2k}\, g_{2k}$:

$$\mathsf{MW}_2^2(f_1, f_2) = \min_{\mathbf{w} \in \Pi(\pi_1, \pi_2)} \sum_{k,l} w_{kl}\, \mathsf{W}_2^2(g_{1k}, g_{2l})$$

- **L_2 distance**

L_2 scalar product between two Gaussian distributions g_1 and g_2:

$$< g_1, g_2 >= \mathcal{N}(\boldsymbol{\mu}_1; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$$

L_2 distance between two Gaussian mixtures f_1 and f_2:

$$\mathsf{L}_2^2(f_1, f_2) = \sum_{k,l} \pi_{1k}\pi_{1l} < g_{1k}, g_{1l} > + \sum_{k,l} \pi_{2k}\pi_{2l} < g_{2k}, g_{2l} > -2 \sum_{k,l} \pi_{1k}\pi_{2l} < g_{1k}, g_{2l} >$$

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: $L_\epsilon(\mathbf{y}, \theta) = \int_{\mathcal{Y}} \mathbb{I}_{\{D(\mathbf{y},\mathbf{z})<\epsilon\}} f_{\boldsymbol{\theta}}(\mathbf{z})\, d\mathbf{z}$

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

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

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

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

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

The requirement $\{\mathbf{z} \in \mathcal{Y},\ D(\mathbf{y}, \mathbf{z}) < \epsilon\} \to \{\mathbf{y}\}$ is too strong

- **An equivalent formulation (Bayes' theorem):**

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

  replace $D(\mathbf{y}, \mathbf{z})$ by $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$, D now a distance on densities

- **A new quasi-posterior:** $q_\epsilon(\boldsymbol{\theta}|\mathbf{y}) \propto \int_{\mathcal{Y}} \mathbb{I}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \le \epsilon\}} \pi(\boldsymbol{\theta}|\mathbf{z})\pi(\mathbf{z}) d\mathbf{z}$

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

Intuition of the proof:

when $\epsilon \to 0$ then $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z})) \to 0$, then $\pi(\cdot \mid \mathbf{z}) \to \pi(\cdot \mid \mathbf{y})$ and

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

$\{\mathbf{z} \in \mathcal{Y}, D\left(\pi(\cdot|\mathbf{y}), \pi(\cdot|\mathbf{z})\right) \le \epsilon\} \to \{\mathbf{z} \in \mathcal{Y}, \pi(\,\cdot|\mathbf{z}) = \pi(\,\cdot|\mathbf{y})\}$ is less demanding

In practice: replace the unknown $\pi(\cdot|\mathbf{y})$ by a tractable approximation

Appendix: Theorem 1 $q_\epsilon(\cdot|\mathbf{y}) \to \pi(\cdot|\mathbf{y})$ in TV

Theorem

For every $\epsilon > 0$, *let* $A_\epsilon = \{\mathbf{z} \in \mathcal{Y} : D\left(\pi\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{z}\right)\right) \leq \epsilon\}$

(A1) $\pi\left(\boldsymbol{\theta} \mid \cdot\right)$ *is continuous for all* $\boldsymbol{\theta} \in \Theta$, *and* $\sup_{\boldsymbol{\theta}\in\Theta} \pi\left(\boldsymbol{\theta} \mid \mathbf{y}\right) < \infty$;

(A2) There exists a $\gamma > 0$ *such that* $\sup_{\boldsymbol{\theta}\in\Theta} \sup_{\mathbf{z}\in A_\gamma} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) < \infty$;

(A3) $D\left(\cdot,\cdot\right) : \Pi \times \Pi \to \mathbb{R}_+$ *is a metric on the functional class*

$$\Pi = \left\{\pi\left(\cdot \mid \mathbf{y}\right) : \mathbf{y} \in \mathcal{Y}\right\};$$

(A4) $D\left(\pi\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{z}\right)\right)$ *is continuous, with respect to* $\mathbf{z}$.

Under (A1)–(A4), $q_\epsilon\left(\cdot \mid \mathbf{y}\right)$ *converges in total variation to* $\pi\left(\cdot \mid \mathbf{y}\right)$, *for fixed* $\mathbf{y}$, *as* $\epsilon \to 0$.

Appendix: proof Theorem 1

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \int_{\mathcal{Y}} K_\epsilon(\mathbf{z};\mathbf{y})\, \pi(\boldsymbol{\theta} \mid \mathbf{z})\, d\mathbf{z} \quad \text{with} \quad K_\epsilon(\mathbf{z};\mathbf{y}) \propto \mathbb{1}_{A_\epsilon}(\mathbf{z})\, \pi(\mathbf{z})$$

$$\begin{aligned}
|q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| &\leq \int_{\mathcal{Y}} K_\epsilon(\mathbf{z};\mathbf{y})\, |\pi(\boldsymbol{\theta} \mid \mathbf{z}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})|\, d\mathbf{z} \\
&\leq \sup_{\mathbf{z} \in A_\epsilon} |\pi(\boldsymbol{\theta} \mid \mathbf{z}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| \quad (K_\epsilon(\cdot;\mathbf{y}) \text{ is a pdf}) \\
&= |\pi(\boldsymbol{\theta} \mid \mathbf{z}_\epsilon) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| \text{ for } \mathbf{z}_\epsilon \in A_\epsilon \ \ (\text{by (A1) and } A_\epsilon \text{ compact})
\end{aligned}$$

For each $\epsilon > 0$, $\mathbf{z}_\epsilon \in A_\epsilon$, $\lim_{\epsilon \to 0} \mathbf{z}_\epsilon \in A_0 = \bigcap_{\epsilon \in \mathbb{Q}_+} A_\epsilon$. Then,
$A_0 = \{\mathbf{z} \in \mathcal{Y} \colon D(\pi(\cdot|\mathbf{z}), \pi(\cdot|\mathbf{y})) = 0\} = \{\mathbf{z} \in \mathcal{Y} \colon \pi(\cdot|\mathbf{z}) = \pi(\cdot|\mathbf{y})\}$ (continuity, equality property of D)

Then $\epsilon \to 0$ yields $|\pi(\boldsymbol{\theta} \mid \mathbf{z}_\epsilon) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| \to |\pi(\boldsymbol{\theta} \mid \mathbf{y}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| = 0$ and hence $|q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| \to 0$, for each $\theta \in \Theta$.

By (A2), $\displaystyle \sup_{\boldsymbol{\theta} \in \Theta} q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \sup_{\boldsymbol{\theta} \in \Theta} \int_{\mathcal{Y}} K_\epsilon(\mathbf{z};\mathbf{y})\, \pi(\boldsymbol{\theta} \mid \mathbf{z})\, d\mathbf{z} \leq \sup_{\boldsymbol{\theta} \in \Theta} \sup_{\mathbf{z} \in A_\gamma} \pi(\boldsymbol{\theta} \mid \mathbf{z}) < \infty$

for some γ, so that $\epsilon \leq \gamma$. Finally (bounded convergence theorem),

$$\lim_{\epsilon \to 0} \int_{\Theta} |q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})|\, d\boldsymbol{\theta} = \lim_{\epsilon \to 0} \|q_\epsilon(\cdot \mid \mathbf{y}) - \pi(\cdot \mid \mathbf{y})\|_1 = 0$$

ABC quasi-posterior with surrogate posteriors $\{p^{K,N}(\cdot|\mathbf{y})\colon \mathbf{y} \in \mathcal{Y}, K \in \mathbb{N}, N \in \mathbb{N}\}$

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

Convergence result for a restricted class of target and surrogate distributions:

$\mathcal{X} = \Theta \times \mathcal{Y}$ compact, $\mathcal{H}_{\mathcal{X}} = \{g_{\boldsymbol{\varphi}} : \boldsymbol{\varphi} \in \Psi\}$ a class of distributions, Ψ bounded,

$$a \leq g_{\boldsymbol{\varphi}}(\mathbf{x}) \leq b \text{ and } \sup_{\mathbf{x} \in \mathcal{X}} |\log g_{\boldsymbol{\varphi}}(\mathbf{x}) - \log g_{\boldsymbol{\varphi}'}(\mathbf{x})| \leq B \|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\|_1$$

Target: $\pi(\mathbf{x}) = \int_{\Psi} g_{\boldsymbol{\varphi}}(\mathbf{x})\, G_\pi(d\boldsymbol{\varphi})$

p^K a K-component mixture of distributions from $\mathcal{H}_{\mathcal{X}}$

$\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n = 1:N\}$ generated from π

$\boldsymbol{\phi}^*_{K,N} = \text{argmax}_{\boldsymbol{\phi} \in \Phi} \sum_{n=1}^N \log\left(p^K(\boldsymbol{\theta}_n, \mathbf{y}_n; \boldsymbol{\phi})\right)$ (MLE)

Surrogates: $p^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) = p^K\left(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}\right)$

Convergence of the ABC quasi-posterior [FF et al, Theorem 2]

Under additional "standard" assumptions

the Hellinger distance $D_H\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ converges to 0

- in some measure λ, with respect to $\mathbf{y} \in \mathcal{Y}$
- in probability, with respect to the sample $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n = 1:N\}$

That is, for any $\alpha > 0, \beta > 0$, it holds that

$$\lim_{\epsilon \to 0, K \to \infty, N \to \infty} \Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \geq \beta\right\}\right) \leq \alpha\right) = 1.$$

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

Appendix: Theorem 2

Theorem

Assume the following: $\mathcal{X} = \Theta \times \mathcal{Y}$ *is a compact set and*

(B1) For joint density π*, there exists* G_π *a probability measure on* Ψ *such that, with* $g_{\boldsymbol{\varphi}} \in \mathcal{H}_{\mathcal{X}}$*,*

$$\pi(\mathbf{x}) = \int_\Psi g_{\boldsymbol{\varphi}}(\mathbf{x})\, G_\pi(d\boldsymbol{\varphi});$$

(B2) The true posterior density $\pi(\cdot \mid \cdot)$ *is continuous both with respect to* $\boldsymbol{\theta}$ *and* $\mathbf{y}$*;*

(B3) $D(\cdot,\cdot) : \Pi \times \Pi \to \mathbb{R}_+ \cup \{0\}$ *is a metric on a functional class* Π*, which contains the class*

$$\left\{p^{K,N}(\cdot \mid \mathbf{y}) : \mathbf{y} \in \mathcal{Y}, K \in \mathbb{N}, N \in \mathbb{N}\right\}.$$

In particular, $D\left(p^{K,N}(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{z})\right) = 0$*, if and only if* $p^{K,N}(\cdot \mid \mathbf{y}) = p^{K,N}(\cdot \mid \mathbf{z})$*;*

(B4) For every $\mathbf{y} \in \mathcal{Y}$*,* $\mathbf{z} \mapsto D\left(p^{K,N}(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{z})\right)$ *is a continuous function on* $\mathcal{Y}$*.*

Then, under (B1)–(B4), the Hellinger distance $D_H\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ *converges to* 0 *in some measure* λ*, with respect to* $\mathbf{y} \in \mathcal{Y}$ *and in probability, with respect to the sample* $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$*. That is, for any* $\alpha > 0, \beta > 0$*, it holds that*

$$\lim_{\epsilon \to 0, K \to \infty, N \to \infty} \Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \geq \beta\right\}\right) \leq \alpha\right) = 1.$$

Appendix: sketch of proof Theorem 2

$$q_\epsilon^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) = \int_{\mathcal{Y}} K_\epsilon^{K,N}(\mathbf{z};\mathbf{y})\, \pi(\boldsymbol{\theta} \mid \mathbf{z})\, d\mathbf{z} \;\text{ with }\; K_\epsilon^{K,N}(\mathbf{z};\mathbf{y}) \propto \mathbb{1}_{A_{\epsilon,\mathbf{y}}^{K,N}}(\mathbf{z})\, \pi(\mathbf{z})$$

Relationship between Hellinger and L_1 distances yields:

$$D_H^2\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \leq 2 D_H\left(\pi(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{y})\right)$$

where $\mathbf{z}_{\epsilon,\mathbf{y}}^{K,N} \in B_{\epsilon,\mathbf{y}}^{K,N}$ with $B_{\epsilon,\mathbf{y}}^{K,N} = \operatorname{argmax}_{\mathbf{z} \in A_{\epsilon,\mathbf{y}}^{K,N}} D_1(\pi(\cdot \mid \mathbf{z}), \pi(\cdot \mid \mathbf{y}))$

$\mathbf{z}_{0,\mathbf{y}}^{K,N} = \lim_{\epsilon \to 0} \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}$ and $\mathbf{z}_{0,\mathbf{y}}^{K,N} \in A_{0,\mathbf{y}}^{K,N} = \left\{\mathbf{z} \in \mathcal{Y} : p^{K,N}(\cdot \mid \mathbf{z}) = p^{K,N}(\cdot \mid \mathbf{y})\right\}$

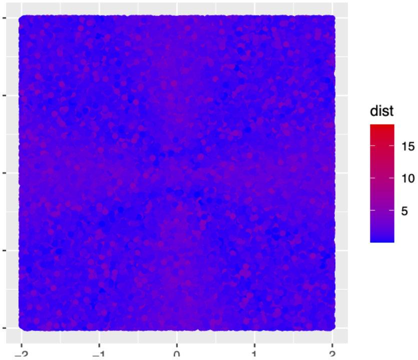
Triangle inequality for D_H:

$$\begin{aligned} D_H\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{y})\right) \leq\; & D_H\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) + D_H\left(\pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}(\cdot \mid \mathbf{y})\right) \\ & + D_H\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \end{aligned}$$

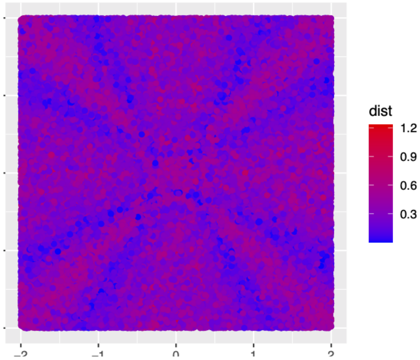
First term in the rhs: goes to 0 as ϵ goes to 0 independently on K, N

Two other terms are similar: use [Rakhlin et al 2005, Corol. 2.2]

Sound source localisation : distance comparison

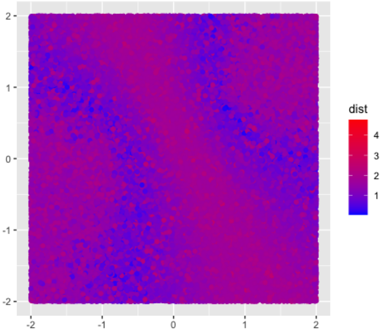


MW_2 distances

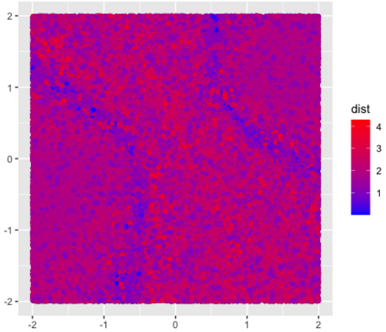


L_2 distances

Sound source localisation : distance comparison



MW_2 distances



L_2 distances

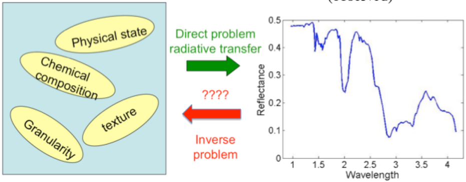
A physical model inversion in planetary science

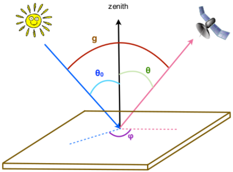
Goal : Study the textural properties of planetary materials

Origin : 1) Remote sensing (Mars surface), 2) Laboratory (analog materials)

Texture and composition parametrized by $\mathbf{x} = (\omega, c, b, \bar{\theta}, B_0, h)$

y : reflectance (observed)

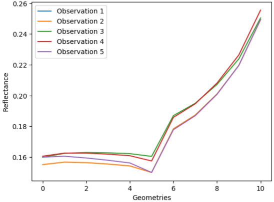




Hapke's radiative transfer model $\mathbf{y} = F(\mathbf{x}) + \varepsilon$

Measurements from 10 geometries

Determination of unknown parameters $(\omega, \bar{\theta}, b, c)$ via reflectance information ($d = 10$ geometries)

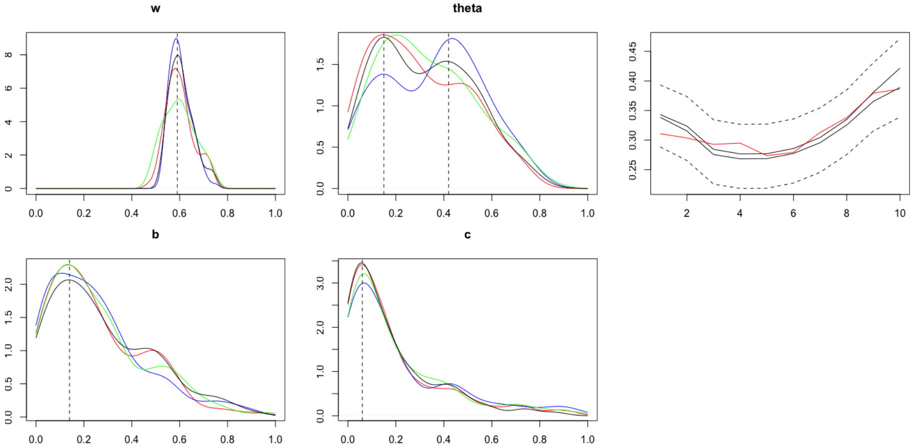


Laboratory observations: Nontronite

GLLiM: $K = 40$, $N = 10^5$; Rejection ABC: $M = 10^5$, ϵ is the 0.1% quantile

1 Nontronite BRDF $\mathbf{y}$: 10 geometries measured (incidence $\theta_0 = 45$, azimuth $\phi = 0$) at 2310nm

$\rightarrow$ Two sets of parameters: $(\omega, \overline{\theta}, b, c) = (0.59, 0.15, 0.14, 0.06)$ and $(0.59, 0.42, 0.14, 0.06)$



Left: GLLiM-E-ABC, GLLiM-L2-ABC, GLLiM-MW2-ABC, Semi-automatic ABC. Right: signal reconstructions

Appendix: other illustration, 10-variate Student likelihood symmetric in μ

$$f_{\boldsymbol{\theta}}(\mathbf{z}) = \mathcal{S}_d(\mathbf{z}; \mu^2 \mathbf{1}_d, \sigma^2 \mathbf{I}_d, \nu)$$

$d = 10$, mean $= (\mu^2 \dots \mu^2)^T$, isotropic scale matrix$= \sigma^2 \mathbf{I}_d$ ($\sigma^2 = 2$), dof (tail) $\nu = 2.1$

Observation $\mathbf{y}$: true $\mu = 1$

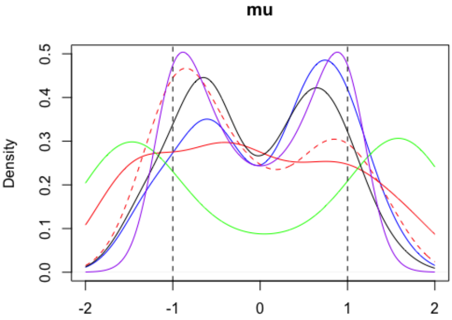
Setting: GLLiM: $K = 10, N = 10^5$; Rejection ABC: $M = 10^5$, $\epsilon = 0.1\%$ (100 values)

True symmetric posterior $\pi(\mu|\mathbf{y})$

GLLiM-E-ABC
GLLiM-EV-ABC (dot)

Semi-automatic ABC

GLLiM-L2-ABC
GLLiM-MW2-ABC



Appendix: other illustration, sum of MA(1) processes

$$y'_t = z_t + \rho z_{t-1}$$
$$y''_t = z'_t - \rho z'_{t-1}$$
$$y_t = y'_t + y''_t$$

$\{z_t\}$ and $\{z'_t\}$ are *i.i.d.* standard normal realizations and ρ is an unknown scalar parameter

$$\rightarrow \mathbf{y} = (y_1, \ldots, y_d)^\top \sim \mathcal{N}(\mathbf{0}_d, 2(\rho^2+1)\mathbf{I}_d)$$

Observation $\mathbf{y}$: $d = 10$, true $\rho = 1$

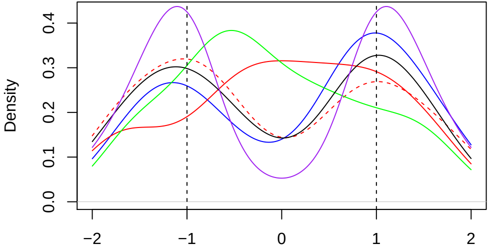
Setting: GLLiM: $K = 20, N = 10^5$; Rejection ABC: $M = 10^5$, $\epsilon = 0.1\%$ (100 selected values)

True symmetric posterior $\pi(\mu|\mathbf{y})$

GLLiM-E-ABC
GLLiM-EV-ABC (dot)

Semi-automatic ABC

GLLiM-L2-ABC
GLLiM-MW2-ABC

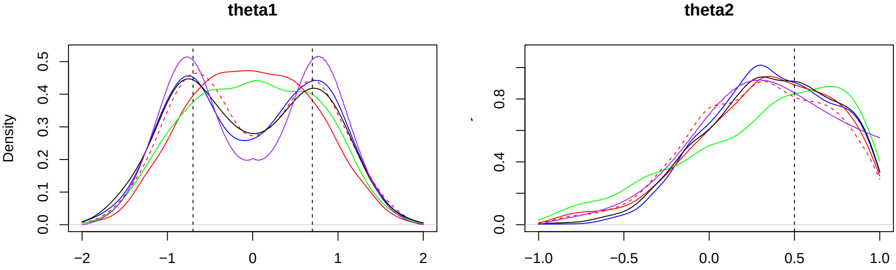


Appendix: other illustration, sum of MA(2)

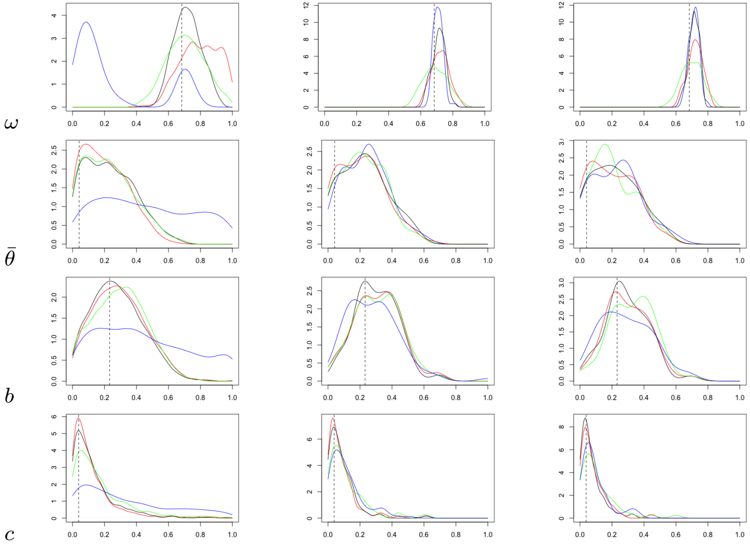
$$
\begin{aligned}
y'_t &= z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2} \\
y''_t &= z'_t - \theta_1 z'_{t-1} + \theta_2 z'_{t-2} \\
y_t &= y'_t + y''_t,
\end{aligned}
$$

$K = 80$ and $N = M = 10^5$, ϵ to the 1% distance quantile (samples of of size 1000)

An observation of size $d = 10$ is simulated from $\theta_1 = 1$ and $\theta_2 = 0.6$



Appendix: synthetic data from the Hapke model



(a) 1% (1000 samples) (b) 0.1% (100 samples) (c) 0.05% (50 samples)