Supporting Information
Approximate Bayesian computation with surrogate posteriors

F. Forbes, H. D. Nguyen, T. Nguyen, J. Arbel

Contents

S1 GLLiM for *i.i.d.* data

The GLLiM implementation of Deleforge et al. (2015) is adapted to account for the fact that for each parameters values, the observations may be available as R *i.i.d.* realizations. The link to the setting where the covariance matrices of the direct model are bloc diagonal is explained. The resulting GLLiM-iid algorithm is detailed. This new procedure can also be useful when dealing with long stationary time series by cutting them into smaller series neglecting dependencies between the sub-series.

S1 .1 Likelihood and posterior approximations with Gaussian mixtures

The Gaussian Locally Linear Mapping (GLLiM) model of Deleforge et al. (2015) is first recalled but note that to match the notation in the manuscript, the notation of Deleforge et al. (2015) has been changed. GLLiM provides probability distributions selected in a family of mixture of Gaussian distributions. An attractive approach for modeling non linear relationships, between some parameters $\boldsymbol{\theta} \in \mathbb{R}^\ell$ and observations $\mathbf{y} \in \mathbb{R}^d$, is to use a mixture of linear models. We assume that each observed $\mathbf{y}$ is the noisy image of parameter $\boldsymbol{\theta}$ obtained from a K-component mixture of affine transformations. This is modeled by introducing a latent variable $z \in \{1, \dots, K\}$ such that

$$\mathbf{y} = \sum_{k=1}^{K} \mathbb{I}_{\{z=k\}} (\tilde{\boldsymbol{A}}_k \boldsymbol{\theta} + \tilde{\boldsymbol{b}}_k + \tilde{\boldsymbol{\epsilon}}_k) \tag{1}$$

where $\mathbb{I}$ is the indicator function, $\tilde{\boldsymbol{A}}_k$ a $d \times \ell$ matrix and $\tilde{\boldsymbol{b}}_k$ a vector of $\mathbb{R}^d$ that define an affine transformation. Variable $\tilde{\boldsymbol{\epsilon}}_k$ corresponds to an error term which is assumed to be zero-mean and not correlated with $\boldsymbol{\theta}$ capturing both the observation noise and the reconstruction error due to the affine approximation. To make the affine transformations local, the latent variable z should also depend on $\boldsymbol{\theta}$.

For the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$ and the likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$ to be easily derived, it is important to control the nature of the joint $p(\mathbf{y}, \boldsymbol{\theta})$. Once a family of tractable joint distributions is chosen, we can look for one that is compatible with (1). In Deleforge et al. (2015) the GLLiM model is derived assuming that the joint distribution is a mixture of Gaussian distributions. Using a subscript G to specify the model, it is assumed that $\tilde{\boldsymbol{\epsilon}}_k \sim \mathcal{N}_d(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_k)$ and that $\boldsymbol{\theta}$ is distributed as a mixture of K Gaussian distributions specified by

$$p_G(\boldsymbol{\theta} | z = k) = \mathcal{N}_\ell(\boldsymbol{\theta}; \tilde{\boldsymbol{c}}_k, \tilde{\boldsymbol{\Gamma}}_k), \tag{2}$$

$$\text{and} \quad p_G(z = k) = \pi_k \, . \tag{3}$$

When informative, we specify the dimension of the Gaussian variable (*e.g.* d) in the notation $\mathcal{N}_d$. The model parameters are then denoted by $\tilde{\boldsymbol{\phi}} = \{\pi_k, \tilde{\boldsymbol{c}}_k, \tilde{\boldsymbol{\Gamma}}_k, \tilde{\boldsymbol{A}}_k, \tilde{\boldsymbol{b}}_k, \tilde{\boldsymbol{\Sigma}}_k\}_{k=1:K}$.

One interesting property of such a parametric model is that both conditional distributions are available in closed form :

$$p_G(\mathbf{y}|\boldsymbol{\theta};\tilde{\boldsymbol{\phi}}) = \sum_{k=1}^{K} \tilde{\eta}_k(\boldsymbol{\theta})\mathcal{N}_d(\mathbf{y};\tilde{\boldsymbol{A}}_k\boldsymbol{\theta}+\tilde{\boldsymbol{b}}_k,\tilde{\boldsymbol{\Sigma}}_k) \quad \text{with } \tilde{\eta}_k(\boldsymbol{\theta}) = \frac{\pi_k\mathcal{N}_\ell(\boldsymbol{\theta};\tilde{\boldsymbol{c}}_k,\tilde{\boldsymbol{\Gamma}}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}_\ell(\boldsymbol{\theta};\tilde{\boldsymbol{c}}_j,\tilde{\boldsymbol{\Gamma}}_j)} \tag{4}$$

$$p_G(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\phi}) = \sum_{k=1}^{K} \eta_k(\mathbf{y})\mathcal{N}_\ell(\boldsymbol{\theta};\boldsymbol{A}_k\mathbf{y}+\boldsymbol{b}_k,\boldsymbol{\Sigma}_k) \quad \text{with } \eta_k(\mathbf{y}) = \frac{\pi_k\mathcal{N}_d(\mathbf{y};\boldsymbol{c}_k,\boldsymbol{\Gamma}_k)}{\sum_{j=1}^{K}\pi_j\mathcal{N}_d(\mathbf{y};\boldsymbol{c}_j,\boldsymbol{\Gamma}_j)}. \tag{5}$$

A different notation $\boldsymbol{\phi}$ is used in (5) but parameters $\boldsymbol{\phi}$ are easily deduced from $\tilde{\boldsymbol{\phi}}$ as follows (the π_k's are unchanged):

$$\begin{aligned}
\boldsymbol{c}_k =& \tilde{\boldsymbol{A}}_k\tilde{\boldsymbol{c}}_k + \tilde{\boldsymbol{b}}_k, \quad \boldsymbol{\Gamma}_k = \tilde{\boldsymbol{\Sigma}}_k + \tilde{\boldsymbol{A}}_k\tilde{\boldsymbol{\Gamma}}_k\tilde{\boldsymbol{A}}_k^\top \\
\boldsymbol{\Sigma}_k =& \left(\tilde{\boldsymbol{\Gamma}}_k^{-1} + \tilde{\boldsymbol{A}}_k^\top\tilde{\boldsymbol{\Sigma}}_k^{-1}\tilde{\boldsymbol{A}}_k\right)^{-1} \\
\boldsymbol{A}_k =& \boldsymbol{\Sigma}_k\tilde{\boldsymbol{A}}_k^\top\tilde{\boldsymbol{\Sigma}}_k^{-1}, \quad \boldsymbol{b}_k = \boldsymbol{\Sigma}_k\left(\tilde{\boldsymbol{\Gamma}}_k^{-1}\tilde{\boldsymbol{c}}_k - \tilde{\boldsymbol{A}}_k^\top\tilde{\boldsymbol{\Sigma}}_k^{-1}\tilde{\boldsymbol{b}}_k\right) .
\end{aligned} \tag{6}$$

The expressions above depend on the value of the parameters $\tilde{\boldsymbol{\phi}}$ that needs to be specified. In Deleforge et al. (2015), parameters $\tilde{\boldsymbol{\phi}}$ are estimated using a maximum likelihood principle with an EM algorithm applied to a learning set of N couples $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$. Once estimated, the parameters lead to an analytical expression of the form (5) denoted by $p_G(\boldsymbol{\theta}|\mathbf{y};\boldsymbol{\phi}^*_{K,N})$, which is a mixture of Gaussian distributions and can be seen as a parametric mapping from $\mathbf{y}$ values to the pdfs on $\boldsymbol{\theta}$. $\boldsymbol{\phi}^*_{K,N}$ can be kept the same for all conditional distributions and does not need to be re-estimated for each new $\boldsymbol{\theta}$ or $\mathbf{y}$ to be inverted.

In practice when d is much larger than ℓ, it is more efficient to estimate $\tilde{\boldsymbol{\phi}}$ from the available data $\mathcal{D}_N$ to then deduce $\boldsymbol{\phi}^*_{K,N}$ and subsequently the conditional distribution of interest (5). The reason is that the size of $\tilde{\boldsymbol{\phi}}$ can be significantly reduced by choosing constraints on matrices $\tilde{\boldsymbol{\Sigma}}_k$ without oversimplifying the target conditional (5). The number of parameters depends on the exact variant learned but is in $\mathcal{O}(dK\ell)$. Typically, diagonal covariance matrices $\tilde{\boldsymbol{\Sigma}}_k$ can be used with a drastic gain. More specifically for the case of diagonal covariances $\tilde{\boldsymbol{\Sigma}}_k$, the number of parameters is $K-1+K(\ell+\ell(\ell+1)/2+d\ell+2d)$ which for $K=100$, $\ell=4$ and $d=10$ leads to 7499 parameters and to 61499 parameters if $d=100$. In addition to the diagonal case, other constraints are implemented in Deleforge et al. (2015), e.g. isotropic, full (no constraint), or equal across k, $\tilde{\boldsymbol{\Sigma}}_k$'s. All details are provided in Deleforge et al. (2015).

In this work, we aim at adapting the GLLiM model and inference to the case of *i.i.d.* observations. It requires the use of another type of constraint, not treated in Deleforge et al. (2015), that induces a bloc diagonal shape for the $\tilde{\boldsymbol{\Sigma}}_k$'s. In the next section we recall the main EM algorithm steps and explain how to modify them to account for this new constraint.

S1 .1.1 GLLiM model parameter estimation

The main updating steps are recalled below.

E-step. The E-step consists in updating the assignments probabilities of each pair $(\boldsymbol{\theta}_n, \mathbf{y}_n)$ to each of the K components, namely for each $k \in [K]$ and $n \in [N]$,

$$r_{nk} \propto \pi_k \, \mathcal{N}_d(\mathbf{y}_n; \tilde{\boldsymbol{A}}_k \boldsymbol{\theta}_n + \tilde{\boldsymbol{b}}_k, \tilde{\boldsymbol{\Sigma}}_k) \, \mathcal{N}_\ell(\boldsymbol{\theta}_n; \tilde{\boldsymbol{c}}_k, \tilde{\boldsymbol{\Gamma}}_k). \tag{7}$$

M-step. Denoting $r_k = \sum_{n=1}^N r_{nk}$, the M-step consists of updating the parameters and decomposes in 3 steps updating successively the π_k's, the $\tilde{\boldsymbol{c}}_k, \tilde{\boldsymbol{\Gamma}}_k$'s and the $\tilde{\boldsymbol{A}}_k, \tilde{\boldsymbol{b}}_k, \tilde{\boldsymbol{\Sigma}}_k$'s.

$$\begin{aligned}
\pi_k &= \frac{r_k}{N} && (8)\\
\tilde{\boldsymbol{c}}_k &= \frac{1}{r_k} \sum_{n=1}^N r_{nk} \boldsymbol{\theta}_n && (9)\\
\tilde{\boldsymbol{\Gamma}}_k &= \frac{1}{r_k} \sum_{n=1}^N r_{nk} (\boldsymbol{\theta}_n - \tilde{\boldsymbol{c}}_k)(\boldsymbol{\theta}_n - \tilde{\boldsymbol{c}}_k)^T && (10)\\
\tilde{\boldsymbol{A}}_k &= \mathbf{Y}_k \boldsymbol{T}_k^T (\boldsymbol{T}_k \boldsymbol{T}_k^T)^{-1} && (11)\\
\tilde{\boldsymbol{b}}_k &= \bar{\mathbf{y}}_k - \tilde{\boldsymbol{A}}_k \tilde{\boldsymbol{c}}_k && (12)\\
\tilde{\boldsymbol{\Sigma}}_k &= \frac{1}{r_k} \sum_{n=1}^N r_{nk} (\mathbf{y}_n - \tilde{\boldsymbol{A}}_k \boldsymbol{\theta}_n - \tilde{\boldsymbol{b}}_k)(\mathbf{y}_n - \tilde{\boldsymbol{A}}_k \boldsymbol{\theta}_n - \tilde{\boldsymbol{b}}_k)^T. && (13)
\end{aligned}$$

The updating of $\tilde{\boldsymbol{A}}_k$ and $\tilde{\boldsymbol{b}}_k$ requires in addition the following quantities depending on the r_{nk}'s and the data set,

$$\begin{aligned}
\bar{\mathbf{y}}_k &= \frac{1}{r_k} \sum_{n=1}^N r_{nk} \mathbf{y}_n \\
\boldsymbol{T}_k &= \frac{1}{\sqrt{r_k}} [\sqrt{r_{1k}}(\boldsymbol{\theta}_1 - \tilde{\boldsymbol{c}}_k) \dots \sqrt{r_{Nk}}(\boldsymbol{\theta}_N - \tilde{\boldsymbol{c}}_k)] \\
\mathbf{Y}_k &= \frac{1}{\sqrt{r_k}} [\sqrt{r_{1k}}(\mathbf{y}_1 - \bar{\mathbf{y}}_k) \dots \sqrt{r_{Nk}}(\mathbf{y}_N - \bar{\mathbf{y}}_k)].
\end{aligned}$$

We now explain how these steps are modified in the *i.i.d.* case.

S1 .2 GLLiM-iid

S1 .2.1 Estimation with bloc constraints

In this section, we consider the case where for a given parameter $\boldsymbol{\theta}$, a sample of R observations $\{\mathbf{y}^1, \dots, \mathbf{y}^R\}$ is generated independently from the likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$. Therefore, we

have $p(\mathbf{y}^1, \ldots, \mathbf{y}^R \mid \boldsymbol{\theta}) = \prod_{r=1}^R p(\mathbf{y}^r \mid \boldsymbol{\theta})$ and we are interested in computing the posterior $p(\boldsymbol{\theta} \mid \mathbf{y}^1, \ldots, \mathbf{y}^R)$. If $R = 1$ we recover the setting handled by standard GLLiM. If $R > 1$, we define a new model and procedure referred to as GLLiM-iid as follows. Note that a key point for our GLLiM-ABC procedure is that the posterior $p(\boldsymbol{\theta} \mid \mathbf{y}^1, \ldots, \mathbf{y}^R)$ still be approximated by a mixture.

Considering the joint $p(\mathbf{y}^1, \ldots, \mathbf{y}^R, \boldsymbol{\theta}) = p(\mathbf{y}^1, \ldots, \mathbf{y}^R \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) = p(\boldsymbol{\theta})\, \prod_{r=1}^R p(\mathbf{y}^r \mid \boldsymbol{\theta})$, we approximate it by using for $p(\boldsymbol{\theta})$ the same mixture model as in standard GLLiM in (2), (3) and for $p(\mathbf{y}^1, \ldots, \mathbf{y}^R \mid \boldsymbol{\theta})$ we assume also a mixture form (we use a different notation q_G to emphasize that it is a particular case of p_G:

$$q_G(\mathbf{y}^1, \ldots, \mathbf{y}^R \mid \boldsymbol{\theta}) \;=\; \sum_{k=1}^K \left(\tilde{\eta}_k(\boldsymbol{\theta}) \prod_{r=1}^R \mathcal{N}_d(\mathbf{y}_r; \tilde{\boldsymbol{A}}_k \boldsymbol{\theta} + \tilde{\boldsymbol{b}}_k, \tilde{\boldsymbol{\Sigma}}_k) \right) . \tag{14}$$

The product in the rhs corresponds to a dR-dimensional Gaussian density with R independent and identical components all of dimension d. This corresponds to a specific constrained GLLiM model where the dimension d has changed to dR and ℓ remains the same. We can then compute the MLE via EM for this constrained parameters setting. The expressions for $\pi_k, \tilde{\boldsymbol{c}}_k, \tilde{\boldsymbol{\Gamma}}_k$ remain the same, while the expressions for $\tilde{\boldsymbol{A}}_k, \tilde{\boldsymbol{b}}_k, \tilde{\boldsymbol{\Sigma}}_k$ are changed using $N \times R$ data points instead of just N. All expressions use a formula for the r_{nk} (the responsibilities) that needs to be modified. This EM algorithm is detailed below.

Observations are now made of R *i.i.d.* vectors of size d and are denoted by $\hat{\mathbf{y}}$ with $\hat{\mathbf{y}} = [\mathbf{y}^1, \ldots, \mathbf{y}^R]^T$. The same expressions (4) and (5) can be used with now parameters $\tilde{\boldsymbol{\phi}}$ denoted by $\hat{\boldsymbol{\phi}}$ and $\hat{\boldsymbol{\Sigma}}_k$ of dimension $dR \times dR$, $\hat{\boldsymbol{A}}_k$ of dimension $dR \times \ell$, $\hat{\boldsymbol{b}}_k$ of length dR, while the dimensions of $\hat{\boldsymbol{c}}_k$ and $\hat{\boldsymbol{\Gamma}}_k$ do not change. Inference could be carried out with the EM described in Section S1 .1.1 with these new dimensions but that would not take into account that the $\mathbf{y}^r$'s are *i.i.d.*. Thus, we propose to add the following constraints, assuming that $\hat{\boldsymbol{\Sigma}}_k$ is a bloc diagonal matrix made of R blocs all equal to $d \times d$ $\tilde{\boldsymbol{\Sigma}}_k$, $\hat{\boldsymbol{A}}_k$ is made of R blocs all equal to $\tilde{\boldsymbol{A}}_k$ of size $d \times \ell$ and $\hat{\boldsymbol{b}}_k$ is a vector of R concatenated vectors all equal to a vector $\tilde{\boldsymbol{b}}_k$ of length d. Note that this is similar to consider a GLLiM model with an additional constraint on the $\hat{\boldsymbol{\Sigma}}_k$'s, namely a bloc diagonal structure except that in this later case no constraint would be assumed on $\hat{\boldsymbol{A}}_k$ and $\hat{\boldsymbol{b}}_k$.

It follows that $q_G(\hat{\mathbf{y}}|\boldsymbol{\theta}, z = k) = \prod_{r=1}^R q_G(\mathbf{y}^r|\boldsymbol{\theta}, z = k)$. As already mentioned, this is not modelling the likelihood as a product but it consists in assuming instead that given $z = k$, so region-wise, the $\mathbf{y}^r$'s are *i.i.d.*

The E-step becomes,

$$r_{nk} \;\propto\; \pi_k\, \mathcal{N}_\ell(\boldsymbol{\theta}_n; \tilde{\boldsymbol{c}}_k, \tilde{\boldsymbol{\Gamma}}_k) \prod_{r=1}^R \mathcal{N}_d(\mathbf{y}_n^r; \tilde{\boldsymbol{A}}_k \boldsymbol{\theta}_n + \tilde{\boldsymbol{b}}_k, \tilde{\boldsymbol{\Sigma}}_k). \tag{15}$$

For the M-step, the expressions for $r_k, \pi_k, \tilde{\boldsymbol{c}}_k, \tilde{\boldsymbol{\Gamma}}_k$ are the same as before (8-10). The rest

of the parameters is modified as follows,

$$\tilde{\boldsymbol{A}}_k = \left(\frac{1}{R}\sum_{r=1}^{R}\mathbf{Y}_k^r\right)\boldsymbol{T}_k^T(\boldsymbol{T}_k\boldsymbol{T}_k^T)^{-1} \tag{16}$$

$$\tilde{\boldsymbol{b}}_k = \frac{1}{R}\sum_{r=1}^{R}\bar{\mathbf{y}}_k^r - \tilde{\boldsymbol{A}}_k\tilde{\boldsymbol{c}}_k \tag{17}$$

$$\tilde{\boldsymbol{\Sigma}}_k = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{r_k}\sum_{n=1}^{N}r_{nk}(\mathbf{y}_n^r - \tilde{\boldsymbol{A}}_k\boldsymbol{\theta}_n - \tilde{\boldsymbol{b}}_k)(\mathbf{y}_n^r - \tilde{\boldsymbol{A}}_k\boldsymbol{\theta}_n - \tilde{\boldsymbol{b}}_k)^T\right) \tag{18}$$

$$= \left(\frac{1}{R}\sum_{r=1}^{R}\right)\bar{\mathbf{y}}_k^r\bar{\mathbf{y}}_k^{rT} - (\tilde{\boldsymbol{A}}_k\boldsymbol{\theta}_n + \tilde{\boldsymbol{b}}_k)(\tilde{\boldsymbol{A}}_k\boldsymbol{\theta}_n + \tilde{\boldsymbol{b}}_k)^T, \tag{19}$$

with

$$\begin{aligned}
\bar{\mathbf{y}}_k^r &= \frac{1}{r_k}\sum_{n=1}^{N}r_{nk}\mathbf{y}_n^r \\
\bar{\mathbf{y}}_k &= \frac{1}{R}\sum_{r=1}^{R}\bar{\mathbf{y}}_k^r \\
\boldsymbol{T}_k &= \frac{1}{\sqrt{r_k}}[\sqrt{r_{1k}}(\boldsymbol{\theta}_1 - \tilde{\boldsymbol{c}}_k)\ldots\sqrt{r_{Nk}}(\boldsymbol{\theta}_N - \tilde{\boldsymbol{c}}_k)] \\
\mathbf{Y}_k^r &= \frac{1}{\sqrt{r_k}}[\sqrt{r_{1k}}(\mathbf{y}_1^r - \bar{\mathbf{y}}_k)\ldots\sqrt{r_{Nk}}(\mathbf{y}_N^r - \bar{\mathbf{y}}_k)].
\end{aligned}$$

Note the use of $\bar{\mathbf{y}}_k$ and not $\bar{\mathbf{y}}_k^r$ in the last expression.

S1 .2.2 Surrogate posteriors in the *i.i.d.* case

The conditional distributions expressions follow from applying the constraint in (4) and (5). Denoting by $\hat{\mathbf{y}}$ the column vector made of the concatenated $\mathbf{y}^1, \ldots, \mathbf{y}^R$,

$$q_G(\boldsymbol{\theta} \mid \mathbf{y}^1, \ldots, \mathbf{y}^R) = \sum_{k=1}^{K}\eta_k^*(\mathbf{y}^1, \ldots, \mathbf{y}^R)\,\mathcal{N}_\ell(\boldsymbol{\theta}; \hat{\boldsymbol{A}}_k^*\hat{\mathbf{y}} + \hat{\boldsymbol{b}}_k^*, \hat{\boldsymbol{\Sigma}}_k^*), \tag{20}$$

where the various parameters and expressions involved are specified below with respect to $\tilde{\boldsymbol{\phi}}$ estimated via the EM algorithm described before:

$$\hat{\boldsymbol{\Sigma}}_k^* = \left(\tilde{\boldsymbol{\Gamma}}_k^{-1} + R\ \tilde{\boldsymbol{A}}_k^T \tilde{\boldsymbol{\Sigma}}_k^{-1} \tilde{\boldsymbol{A}}_k\right)^{-1} \tag{21}$$

$$\hat{\boldsymbol{A}}_k^* \hat{\mathbf{y}} + \hat{\boldsymbol{b}}_k^* = \hat{\boldsymbol{\Sigma}}_k^* \left(\tilde{\boldsymbol{A}}_k^T \tilde{\boldsymbol{\Sigma}}_k^{-1} (\sum_{r=1}^{R} \mathbf{y}^r) + \tilde{\boldsymbol{\Gamma}}_k^{-1} \tilde{\boldsymbol{c}}_k - R\ \tilde{\boldsymbol{A}}_k^T \tilde{\boldsymbol{\Sigma}}_k^{-1} \tilde{\boldsymbol{b}}_k \right) \tag{22}$$

$$\eta_k^*(\mathbf{y}^1, \ldots, \mathbf{y}^R) \propto \pi_k\, \mathcal{N}_{dR}(\tilde{\mathbf{y}}; \boldsymbol{m}_k, \boldsymbol{V}_k), \tag{23}$$

where $\boldsymbol{m}_k$ is a vector made of R concatenated d-dimensional vectors all equal to $\tilde{\boldsymbol{A}}_k \tilde{\boldsymbol{c}}_k + \tilde{\boldsymbol{b}}_k$ and $\boldsymbol{V}_k$ is a matrix made of $R \times R$ blocs of size $d \times d$ which is the sum of a bloc diagonal matrix with all diagonal blocs equal to $\tilde{\boldsymbol{\Sigma}}_k$ and of a matrix made of constant blocs all equal to $\tilde{\boldsymbol{A}}_k \tilde{\boldsymbol{\Gamma}}_k \tilde{\boldsymbol{A}}_k^T$.

As dR can be large, *e.g.* 1000, the computation of η_k^* can be numerically problematic. However the quadratic forms and the determinants involved can simplify using the Woodbury formula and the matrix determinant lemma. Let $S_k = (\hat{\mathbf{y}} - \boldsymbol{m}_k)^T \boldsymbol{V}_k^{-1} (\hat{\mathbf{y}} - \boldsymbol{m}_k)$, the Woodbury formula leads to:

$$\begin{aligned} S_k &= \sum_{r=1}^{R} (\mathbf{y}^r - \tilde{\boldsymbol{A}}_k \tilde{\boldsymbol{c}}_k - \tilde{\boldsymbol{b}}_k)^T \tilde{\boldsymbol{\Sigma}}_k^{-1} (\mathbf{y}^r - \tilde{\boldsymbol{A}}_k \tilde{\boldsymbol{c}}_k - \tilde{\boldsymbol{b}}_k) \\ &\quad - \left(\sum_{r=1}^{R} \mathbf{y}^r - R \tilde{\boldsymbol{A}}_k \tilde{\boldsymbol{c}}_k - R \tilde{\boldsymbol{b}}_k \right)^T \left(\tilde{\boldsymbol{\Sigma}}_k^{-1} \tilde{\boldsymbol{A}}_k \hat{\boldsymbol{\Sigma}}_k^* \tilde{\boldsymbol{A}}_k^T \tilde{\boldsymbol{\Sigma}}_k^{-1} \right) \left(\sum_{r=1}^{R} \mathbf{y}^r - R \tilde{\boldsymbol{A}}_k \tilde{\boldsymbol{c}}_k - R \tilde{\boldsymbol{b}}_k \right) . \end{aligned}$$

Similarly for the determinant of $\boldsymbol{V}_k$ we get:

$$|\boldsymbol{V}_k| = |\tilde{\boldsymbol{\Sigma}}_k|^R \times |\boldsymbol{I} + R\ \tilde{\boldsymbol{\Gamma}}_k^{1/2} \tilde{\boldsymbol{A}}_k^T \tilde{\boldsymbol{\Sigma}}_k^{-1} \tilde{\boldsymbol{A}}_k \tilde{\boldsymbol{\Gamma}}_k^{1/2}| \,,$$

or equivalently

$$|\boldsymbol{V}_k| = |\tilde{\boldsymbol{\Sigma}}_k|^R \times |\boldsymbol{I} + R\ \tilde{\boldsymbol{\Gamma}}_k \tilde{\boldsymbol{A}}_k^T \tilde{\boldsymbol{\Sigma}}_k^{-1} \tilde{\boldsymbol{A}}_k|.$$

So that

$$\log \eta_k^* = \log \pi_k - 0.5 S_k - 0.5 \log |\boldsymbol{V}_k| + C \,.$$

The estimated $\tilde{\boldsymbol{\phi}}_R$ parameters can then be used to specify (20).

S2 Distances between Gaussian mixtures

S2 .1 Optimal transport-based distance between Gaussian mixtures

Delon and Desolneux (2020); Chen et al. (2019) have introduced a distance specifically designed for Gaussian mixtures based on the Wasserstein distance. In an optimal transport context, by restricting the possible coupling measures (*i.e.*, the optimal transport plan) to

a Gaussian mixture, they propose a discrete formulation for this distance. This makes it tractable and suitable for high dimensional problems, while in general using the standard Wasserstein distance between mixtures is problematic. Delon and Desolneux (2020) refer to the proposed new distance as MW_2, for *Mixture Wasserstein*.

The MW_2 definition makes first use of the tractability of the Wasserstein distance between two Gaussians for a quadratic cost. The standard quadratic cost Wasserstein distance between two Gaussian pdfs $g_1(\cdot) = \mathcal{N}(\cdot\,; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $g_2(\cdot) = \mathcal{N}(\cdot\,; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ is (see Delon and Desolneux 2020),

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

Section 4 of Delon and Desolneux (2020) shows that the MW_2 distance between two mixtures can be computed by solving the following discrete optimization problem. Let $f_1 = \sum_{k=1}^{K_1} \pi_{1k}\, g_{1k}$ and by $f_2 = \sum_{k=1}^{K_2} \pi_{2k}\, g_{2k}$ be two Gaussian mixtures. Then,

$$\mathrm{MW}_2^2(f_1, f_2) = \min_{\mathbf{w} \in \Pi(\pi_1, \pi_2)} \sum_{k,l} w_{kl}\, \mathrm{W}_2^2(g_{1k}, g_{2l})\ , \tag{24}$$

where π_1 and π_2 are the discrete distributions on the simplex defined by the respective weights of the mixtures and $\Pi(\pi_1, \pi_2)$ is the set of discrete joint distributions $\mathbf{w} = (w_{kl}, k \in [K_1], l \in [K_2])$, whose marginals are π_1 and π_2. Finding the minimizer $\mathbf{w}^*$ of (24) boils down to solving a simple discrete optimal transport problem, where the entries of the $K_1 \times K_2$ dimensional cost matrix are the $\mathrm{W}_2^2(g_{1k}, g_{2l})$ quantities.

As implicitly suggested above, MW_2 is indeed a distance on the space of Gaussian mixtures; see Delon and Desolneux (2020). In particular, for two Gaussian mixtures f_1 and f_2, MW_2 satisfies the equality property according to which $\mathrm{MW}_2(f_1, f_2) = 0$ implies that $f_1 = f_2$. In our experiments, the MW_2 distances were computed using the **transport** R package (Schuhmacher et al., 2020).

S2 .2 L_2 distance between Gaussian mixtures

The L_2 distance between two Gaussian mixtures is also closed form. Denote by $f_1 = \sum_{k=1}^{K_1} \pi_{1k}\, g_{1k}$ and $f_2 = \sum_{k=1}^{K_2} \pi_{2k}\, g_{2k}$ two Gaussian mixtures,

$$\mathrm{L}_2^2(f_1, f_2) = \sum_{k,l} \pi_{1k}\pi_{1l}\langle g_{1k}, g_{1l}\rangle + \sum_{k,l} \pi_{2k}\pi_{2l}\langle g_{2k}, g_{2l}\rangle - 2\sum_{k,l} \pi_{1k}\pi_{2l}\langle g_{1k}, g_{2l}\rangle, \tag{25}$$

where $\langle\cdot,\cdot\rangle$ denotes the L_2 scalar product, which is closed form for two Gaussian distributions g_1 and g_2 and given by $\langle g_1, g_2\rangle = \mathcal{N}(\boldsymbol{\mu}_1; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)$. The L_2 distance can be evaluated in $\mathcal{O}(K_1K_2)$ time. We do not discuss the different properties of the various possible distances but the distance choice has a potential impact on the associated GLLiM-D-ABC procedure. This impact is illustrated in the experimental Section S4 .

S3 Proofs

S3 .1 Proof of Theorem 1

We follow steps similar to the proof of Proposition 2 in Bernton et al. (2019). The ABC quasi-posterior can be written as

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \int_{\mathcal{Y}} K_\epsilon(\mathbf{z}; \mathbf{y})\, \pi(\boldsymbol{\theta} \mid \mathbf{z})\, d\mathbf{z},$$

where $K_\epsilon(\mathbf{z}; \mathbf{y}) \propto \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}), \pi(\cdot|\mathbf{z})) \leq \epsilon\}}\, \pi(\mathbf{z})$ denotes the density evaluated at some $\mathbf{z}$ of the prior truncated to A_ϵ. $K_\epsilon(\cdot; \mathbf{y})$ is a probability density function (pdf) in $\mathbf{z} \in \mathcal{Y}$ with compact support $A_\epsilon \subset \mathcal{Y}$ by definition of A_ϵ and (A4). It follows that

$$\begin{aligned} |q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| &\leq \int_{\mathcal{Y}} K_\epsilon(\mathbf{z}; \mathbf{y})\, |\pi(\boldsymbol{\theta} \mid \mathbf{z}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})|\, d\mathbf{z} \\ &\leq \sup_{\mathbf{z} \in A_\epsilon} |\pi(\boldsymbol{\theta} \mid \mathbf{z}) - \pi(\boldsymbol{\theta} \mid \mathbf{y})| \\ &= |\pi(\boldsymbol{\theta} \mid \mathbf{z}_\epsilon) - \pi(\boldsymbol{\theta} \mid \mathbf{y})|, \end{aligned}$$

for some $\mathbf{z}_\epsilon \in A_\epsilon$, where the second inequality is due to the fact that $K_\epsilon(\cdot; \mathbf{y})$ is a pdf, and the last equality is due to (A1) and the compacity of A_ϵ.

Since for each $\epsilon > 0$, $\mathbf{z}_\epsilon \in A_\epsilon$, we have $\lim_{\epsilon \to 0} \mathbf{z}_\epsilon \in A_0$, where $A_0 = \bigcap_{\epsilon \in \mathbb{Q}_+} A_\epsilon$. Then, using that by continuity of D, $A_0 = \{\mathbf{z} \in \mathcal{Y} : D(\pi(\cdot \mid \mathbf{z}), \pi(\cdot \mid \mathbf{y})) = 0\}$, it follows from the equality property of D, that $A_0 = \{\mathbf{z} \in \mathcal{Y} : \pi(\cdot \mid \mathbf{z}) = \pi(\cdot \mid \mathbf{y})\}$. Taking the limit $\epsilon \to 0$ 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), we have

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

for some γ, so that $\epsilon \leq \gamma$. Finally, by the bounded convergence theorem, we have

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

S3 .2 Proof of Theorem 2

We now provide a detailed proof of Theorem 2. Given any $\alpha > 0, \beta > 0$, we claim that

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

or equivalently, for any $\alpha > 0, \beta > 0, \gamma > 0$, we wish to find $\epsilon(\alpha,\beta,\gamma) > 0$, $K(\alpha,\beta,\gamma) \in \mathbb{N}^*$, and $N(\alpha,\beta,\gamma) \in \mathbb{N}^*$ so that for all $\epsilon < \epsilon(\alpha,\beta,\gamma), K \geq K(\alpha,\beta,\gamma), N \geq N(\alpha,\beta,\gamma)$:

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \geq \beta\right\}\right) > \alpha\right) \leq \gamma. \tag{26}$$

To prove (26), we first recall that we can rewrite $q_\epsilon^{K,N}$ as follows, for all $\boldsymbol{\theta} \in \Theta, \mathbf{y} \in \mathcal{Y}$,

$$\begin{aligned} q_\epsilon^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) &= \int_{\mathcal{Y}} K_\epsilon^{K,N}(\mathbf{z};\mathbf{y})\, \pi(\boldsymbol{\theta} \mid \mathbf{z})\, d\mathbf{z}, \\ K_\epsilon^{K,N}(\mathbf{z};\mathbf{y}) &= \frac{\mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}), p^{K,N}(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\mathbf{z})}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}), p^{K,N}(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\mathbf{z})\, d\mathbf{z}}, \end{aligned} \tag{27}$$

where $K_\epsilon^{K,N}(\cdot;\mathbf{y})$ is a pdf on $\mathbf{z} \in \mathcal{Y}$ with compact support $A_{\epsilon,\mathbf{y}}^{K,N} \subset \mathcal{Y}$ by definition of $A_{\epsilon,\mathbf{y}}^{K,N}$ and (B4).

The Hellinger distance D_{H}, between two densities f and g in appropriate spaces, is related to the L_1 distance D_1 as follows, see Zeevi and Meir (1997, Lemma 1),

$$\left(\frac{1}{2} D_1(f,g)\right)^2 \leq D_H^2(f,g) \leq D_1(f,g). \tag{28}$$

Applying successively the right-hand-side of (28), the definition of $q_\epsilon^{K,N}$ and the fact that $K_\epsilon^{K,N}(\cdot;\mathbf{y})$ is a pdf, we can write

$$\begin{aligned} D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) &\leq D_1\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \\ &= \int_\Theta \left| q_\epsilon^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) - \pi(\boldsymbol{\theta} \mid \mathbf{y}) \right| d\lambda(\boldsymbol{\theta}) \\ &\leq \int_\Theta \int_{\mathcal{Y}} K_\epsilon^{K,N}(\mathbf{z};\mathbf{y}) \left| \pi(\boldsymbol{\theta} \mid \mathbf{z}) - \pi(\boldsymbol{\theta} \mid \mathbf{y}) \right| d\lambda(\mathbf{z})\, d\lambda(\boldsymbol{\theta}) \\ &= \int_{\mathcal{Y}} K_\epsilon^{K,N}(\mathbf{z};\mathbf{y}) \int_\Theta \left| \pi(\boldsymbol{\theta} \mid \mathbf{z}) - \pi(\boldsymbol{\theta} \mid \mathbf{y}) \right| d\lambda(\boldsymbol{\theta})\, d\lambda(\mathbf{z}) \\ &\leq \sup_{\mathbf{z} \in A_{\epsilon,\mathbf{y}}^{K,N}} \int_\Theta \left| \pi(\boldsymbol{\theta} \mid \mathbf{z}) - \pi(\boldsymbol{\theta} \mid \mathbf{y}) \right| d\lambda(\boldsymbol{\theta}) \ . \end{aligned}$$

Then using Makarov and Podkorytov (2013, Corollary 7.1.3) and the continuity of $\pi(\cdot \mid \cdot)$ (B2), it follows that $\mathbf{z} \mapsto D_1(\pi(\cdot \mid \mathbf{z}), \pi(\cdot \mid \mathbf{y}))$ is a continuous function for every $\mathbf{y} \in \mathcal{Y}$. As $A_{\epsilon,\mathbf{y}}^{K,N}$ is compact, since

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

$$\sup_{\mathbf{z} \in A_{\epsilon,\mathbf{y}}^{K,N}} D_1(\pi(\cdot \mid \mathbf{z}), \pi(\cdot \mid \mathbf{y})) = D_1\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{y})\right),$$

and using the left-hand-side of (28), we finally get that

$$D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) \leq 2D_{\mathrm{H}}\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right). \tag{29}$$

Consider the limit point $\mathbf{z}_{0,\mathbf{y}}^{K,N}$ defined as $\mathbf{z}_{0,\mathbf{y}}^{K,N} = \lim_{\epsilon\to 0} \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}$. Since for each $\epsilon > 0$, $\mathbf{z}_{\epsilon,\mathbf{y}}^{K,N} \in A_{\epsilon,\mathbf{y}}^{K,N}$ then $\mathbf{z}_{0,\mathbf{y}}^{K,N} \in A_{0,\mathbf{y}}^{K,N}$, where $A_{0,\mathbf{y}}^{K,N} = \bigcap_{\epsilon\in\mathbb{Q}_+} A_{\epsilon,\mathbf{y}}^{K,N}$. By continuity of D, $A_{0,\mathbf{y}}^{K,N} = \left\{\mathbf{z} \in \mathcal{Y} : D\left(p^{K,N}\left(\cdot \mid \mathbf{z}\right), p^{K,N}\left(\cdot \mid \mathbf{y}\right)\right) = 0\right\}$ and $A_{0,\mathbf{y}}^{K,N} = \left\{\mathbf{z} \in \mathcal{Y} : p^{K,N}\left(\cdot \mid \mathbf{z}\right) = p^{K,N}\left(\cdot \mid \mathbf{y}\right)\right\}$, using (B3).

The distance on the right-hand side of (29) can then be bounded by three terms using the triangle inequality for the Hellinger distance D_{H},

$$\begin{aligned} D_{\mathrm{H}}\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) \leq\ & D_{\mathrm{H}}\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) + D_{\mathrm{H}}\left(\pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}\left(\cdot \mid \mathbf{y}\right)\right) \\ & + D_{\mathrm{H}}\left(p^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right). \end{aligned} \tag{30}$$

The first term on the right-hand side can be made close to 0 as ϵ goes to 0 independently of K and N. The two other terms are of the same nature as the definition of $\mathbf{z}_{0,\mathbf{y}}^{K,N}$ yields $p^{K,N}\left(\cdot \mid \mathbf{y}\right) = p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})$.

Therefore, we first prove that $\lim_{\epsilon\to 0} D_H^2\left(\pi(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) = 0$ pointwise *i.e.* for each $\mathbf{y}$. Indeed, since $\pi(\cdot \mid \cdot)$ is a uniformly continuous function in $(\boldsymbol{\theta}, \mathbf{y})$, given any $\mathbf{y} \in \mathcal{Y}$, $\alpha_1 > 0$, there exists $\delta(\alpha_1) > 0$ such that for all $\mathbf{z}_{0,\mathbf{y}}^{K,N} \in A_{0,\mathbf{y}}^{K,N} \subset \mathcal{Y}$,

$$\sup_{\boldsymbol{\theta}\in\Theta} \left|\pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) - \pi(\boldsymbol{\theta} \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right| \leq \alpha_1, \forall \mathbf{z} \in \mathcal{Y}, \left|\mathbf{z} - \mathbf{z}_{0,\mathbf{y}}^{K,N}\right| < \delta(\alpha_1). \tag{31}$$

Furthermore, since Θ is a subset of a compact set, $\lambda\left(\Theta\right) < \infty$. Hence, by using the fact that $\lim_{\epsilon\to 0} \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N} = \mathbf{z}_{0,\mathbf{y}}^{K,N} \in A_{0,\mathbf{y}}^{K,N}$ pointwise with respect to $\mathbf{y}$ and choosing $\mathbf{z} = \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}$ in (31), we obtain that given any $\mathbf{y} \in \mathcal{Y}$, and $\alpha_1 > 0$, there exists $\delta(\alpha_1) > 0$, and $\epsilon\left(\delta(\alpha_1)\right) > 0$ such that $\forall 0 < \epsilon < \epsilon\left(\delta(\alpha_1)\right)$, $\left|\mathbf{z}_{\epsilon,\mathbf{y}}^{K,N} - \mathbf{z}_{0,\mathbf{y}}^{K,N}\right| < \delta(\alpha_1)$. Using (28) and (31), it follows for any ϵ such that $0 < \epsilon < \epsilon\left(\delta(\alpha_1)\right)$,

$$\begin{aligned} D_H^2\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) &\leq D_1\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \\ &\leq \sup_{\boldsymbol{\theta}\in\Theta} \left|\pi\left(\boldsymbol{\theta} \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right) - \pi(\boldsymbol{\theta} \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right| \lambda\left(\Theta\right) \\ &\leq \alpha_1 \lambda\left(\Theta\right). \end{aligned} \tag{32}$$

Such convergence also holds in measure λ. Given any $\alpha_1 > 0$, $\beta_1 > 0$, there exists $\epsilon\left(\alpha_1, \beta_1\right) > 0$ such that for any $0 < \epsilon < \epsilon\left(\alpha_1, \beta_1\right)$,

$$\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \geq \beta_1\right\}\right) \leq \alpha_1. \tag{33}$$

Then, since (33) is true whatever the value of $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$, sampled from the joint $\pi(\cdot, \cdot)$, it also holds, in probability with respect to the data set, that

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \geq \beta_1\right\}\right) > \alpha_1\right) = 0. \tag{34}$$

Next, we prove that $D_H^2\Big(\pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}(\cdot \mid \mathbf{y})\Big)$, equal to $D_H^2\Big(\pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\Big)$, and $D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ both converge to 0 in measure λ, with respect to $\mathbf{y}$ and in probability with respect to the sample $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$.

We first focus on $D_H^2\left(p^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right)$. Using the monotonicity of the Lebesgue integral and a result from Tsybakov (2008, Lemma 2.4) indicating that the squared Hellinger distance can be bounded by the Kullback–Leibler (KL) divergence, it follows that

$$\int_{\mathcal{Y}} D_H^2\left(p^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) d\lambda(\mathbf{y}) \leq \int_{\mathcal{Y}} \mathrm{KL}\left(\pi(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{y})\right) d\lambda(\mathbf{y}).$$

Then since $\pi(\mathbf{y}) \geq a\lambda(\Theta)$

$$\begin{aligned}\int_{\mathcal{Y}} \mathrm{KL}\left(\pi(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{y})\right) d\lambda(\mathbf{y}) &\leq \frac{1}{a\lambda(\Theta)} \int_{\mathcal{Y}} \pi(\mathbf{y})\ \mathrm{KL}\left(\pi(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{y})\right) d\lambda(\mathbf{y}) \\ &\leq \frac{1}{a\lambda(\Theta)}\ \mathrm{KL}\left(\pi, p^{K,N}\right)\ , \end{aligned} \tag{35}$$

where in the last right-hand side, the Kullback–Leibler divergence is on the joint densities π and $p^{K,N}$ and the inequality is coming from a standard relationship between Kullback–Leibler divergences between joint and conditional distributions, *i.e.*

$$\mathrm{KL}\left(\pi, p^{K,N}\right) = \int_{\mathcal{Y}} \pi(\mathbf{y})\ \mathrm{KL}\left(\pi(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{y})\right) d\lambda(\mathbf{y}) + \int_{\mathcal{Y}} \pi(\mathbf{y})\ \log\left(\frac{\pi(\mathbf{y})}{p^{K,N}(\mathbf{y})}\right)\ d\lambda(\mathbf{y})\ ,$$

with the last integral being a positive Kullback–Leibler divergence. Using Corollary 2.2 in Rakhlin et al. (2005) (see details in Section S3 .3.1), we can show that $\mathrm{KL}\left(\pi, p^{K,N}\right)$ tends to 0 in probability as K and N tends to infinity. It follows that $D_H^2\left(p^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right)$ converges to 0 in L_1 distance with respect to $\mathbf{y}$. Using Tao (2011, 1.5. Modes of convergence), $D_H^2\left(p^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right)$ also converges to 0 in measure λ with respect to $\mathbf{y}$, and in probability with respect to the sample $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ as $K \to \infty, N \to \infty$.
That is, given any $\alpha_2 > 0$, $\beta_2 > 0$, $\gamma_2 > 0$, there exists $K\left(\alpha_2, \beta_2, \gamma_2\right) \in \mathbb{N}^*$, $N\left(\alpha_2, \beta_2, \gamma_2\right) \in \mathbb{N}^*$ such that for any $K \geq K\left(\alpha_2, \beta_2, \gamma_2\right)$, $N \geq N\left(\alpha_2, \beta_2, \gamma_2\right)$,

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y}, D_H^2\left(p^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) \geq \beta_2\right\}\right) > \alpha_2\right) \leq \gamma_2. \tag{36}$$

To show that the same as (36) also holds when replacing $\mathbf{y}$ by $\mathbf{z}_{0,\mathbf{y}}^{K,N}$ in D_H^2, we need to show some measurability property with respect to λ. Lemma 2, together with its proof in Subsection S3 .3.2, guaranties first that the map $\mathbf{y} \mapsto \mathbf{z}_0^{K,N}\left(\mathbf{y}\right) = \mathbf{z}_{0,\mathbf{y}}^{K,N}$ is measurable.

Since $\mathbf{y} \mapsto D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ is a continuous function (using (B4) and Makarov and Podkorytov 2013, Corollary 7.1.3), the measurability of the map implies that $D_H^2\left(p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right)$ is also a measurable function (see Tao 2011, 1.3.2. Measurable functions). Consequently Tao (2011, Lemma 1.3.9 Equivalent notions of measurability) the set $\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \geq \beta_2\right\}$ is a measurable set with respect to λ. In addition by the monotonicity of λ and the defintion of $\mathbf{z}_{0,\mathbf{y}}^{K,N}$, the measure of this set satisfies for any $\beta_2 > 0$,

$$\lambda(\{\mathbf{y} \in \mathcal{Y} : D_H^2(p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})) \geq \beta_2\}) \leq \lambda(\{\mathbf{y} \in \mathcal{Y} : D_H^2(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})) \geq \beta_2\}).$$

Then (36) implies that

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \geq \beta_2\right\}\right) > \alpha_2\right) \leq \gamma_2. \tag{37}$$

Finally, (26) can be deduced from (34), (36) and (37) by choosing $\alpha_1 = \alpha_2 = \alpha/3$, $\beta_1 = \beta_2 = \beta^2/36$, $\gamma_2 = \gamma/2$, $\epsilon(\alpha, \beta, \gamma) = \epsilon(\alpha_1, \beta_1)$, $K(\alpha, \beta, \gamma) = K(\alpha_2, \beta_2, \gamma_2)$ and $N(\alpha, \beta, \gamma) = N(\alpha_2, \beta_2, \gamma_2)$.

S3 .3 Auxiliary results

S3 .3.1 Use of Corollary 2.2 of Rakhlin et al. (2005)

In this section, we claim that under the conditions of Theorem 2, we can prove that $\mathrm{KL}\left(\pi, p^{K,N}\right) \to 0$, in probability as $K \to \infty, N \to \infty$.

To do so we use the following Lemma 1 coming from Rakhlin et al. (2005). Let us recall that $\mathcal{H}_{\mathcal{X}}$ is a parametric family of pdfs on $\mathcal{X}$, $\mathcal{H}_{\mathcal{X}} = \{g_{\boldsymbol{\varphi}}, \boldsymbol{\varphi} \in \Psi\}$. The set of continuous convex combinations associated with $\mathcal{H}_{\mathcal{X}}$ is defined as

$$\mathcal{C} = \mathrm{conv}(\mathcal{H}_{\mathcal{X}}) = \left\{f : f(\mathbf{x}) = \int_{\Psi} g_{\boldsymbol{\varphi}}(\mathbf{x})\, G(d\boldsymbol{\varphi}),\, g_{\boldsymbol{\varphi}} \in \mathcal{H}_{\mathcal{X}}, G \text{ is a probability measure on } \Psi\right\}.$$

We write $\mathrm{KL}(\pi, \mathcal{C}) = \inf_{g \in \mathcal{C}} \mathrm{KL}(\pi, g)$.
The class of K-component mixtures on $\mathcal{H}_{\mathcal{X}}$ is then defined as

$$\mathcal{C}_K = \mathrm{conv}_K(\mathcal{H}_{\mathcal{X}}) = \left\{f : f(\mathbf{x}) = \sum_{k=1}^{K} c_k g_{\boldsymbol{\varphi}_k}(\mathbf{x}),\, c \in \mathbb{S}^{K-1}, g_{\boldsymbol{\varphi}_k} \in \mathcal{H}_{\mathcal{X}}\right\} \tag{38}$$

where $\mathbb{S}^{K-1} = \left\{(c_1, \ldots, c_K) \in \mathbb{R}^K : \sum_{k=1}^{K} c_k = 1, c_k \geq 0, k \in [K]\right\}$.

The result from Rakhlin et al. (2005) is recalled in the following Lemma.

Lemma 1 (Corollary 2.2. from Rakhlin et al. (2005)) *Let $\mathcal{X} = \Theta \times \mathcal{Y}$ be a compact set. Let π be a target density π such that $0 < a \leq \pi(\mathbf{x}) \leq b$, for all $\mathbf{x} \in \mathcal{X}$. Assume that the distributions in $\mathcal{H}_{\mathcal{X}}$ satisfy, for any $\boldsymbol{\varphi}, \boldsymbol{\varphi}' \in \Psi$,*

$$\begin{aligned} &\text{for all } \mathbf{x} \in \mathcal{X},\ 0 < a \leq g_{\boldsymbol{\varphi}}(\mathbf{x}) \leq b \\ &\text{and } \sup_{\mathbf{x} \in \mathcal{X}} \left| \log g_{\boldsymbol{\varphi}}(\mathbf{x}) - \log g_{\boldsymbol{\varphi}'}(\mathbf{x}) \right| \leq B \|\boldsymbol{\varphi} - \boldsymbol{\varphi}'\|_1 \,, \end{aligned}$$

and that the parameter set Ψ is a cube with side length A with a, b, A, B arbitrary positive scalars. Let $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ be realizations from the joint distribution $\pi(\cdot, \cdot)$ and denote by $p^{K,N}$ the K-component mixture MLE in $\mathcal{C}_K$.
Then, with probability at least $1 - \exp(-t)$,

$$\mathrm{KL}\left(\pi, p^{K,N}\right) \leq \mathrm{KL}\left(\pi, \mathcal{C}\right) + \frac{c_1}{K} + \frac{c_2}{\sqrt{N}} + \frac{c_3\sqrt{t}}{\sqrt{N}},$$

where c_1, c_2 and c_3 are positive scalars depending only on a, b, A, B and on the dimension of $\mathcal{X}$ (see Rakhlin et al. (2005) for the exact expressions).

Assumption (B1) in Theorem 2 then implies that $\pi \in \mathcal{C}$ so that $\mathrm{KL}\left(\pi, \mathcal{C}\right) = 0$. Using Lemma 1, it follows that for all $t > 0$, for all $K \in \mathbb{N}^*$, and for all $N \in \mathbb{N}^*$,

$$\Pr\left(\mathrm{KL}\left(\pi, p^{K,N}\right) \leq \frac{c_1}{K} + \frac{c_2}{\sqrt{N}} + \frac{c_3\sqrt{t}}{\sqrt{N}}\right) \geq 1 - \exp\left(-t\right). \tag{39}$$

Choosing $t = N^{1/2}$, (39) becomes

$$1 - \Pr\left(\mathrm{KL}\left(\pi, p^{K,N}\right) \leq \frac{c_1}{K} + \frac{c_2}{\sqrt{N}} + \frac{c_3}{N^{1/4}}\right) \leq \exp\left(-N^{1/2}\right). \tag{40}$$

Therefore, for any $\gamma_1 > 0, \gamma_2 > 0$, there exist $K(\gamma_1, \gamma_2) \in \mathbb{N}^*$, and $N(\gamma_1, \gamma_2) \in \mathbb{N}^*$ so that for all $K \geq K(\gamma_1, \gamma_2)$ and $N \geq N(\gamma_1, \gamma_2)$,

$$\begin{aligned} &\frac{c_1}{K} + \frac{c_2}{\sqrt{N}} + \frac{c_3}{N^{1/4}} \leq \gamma_1, \\ &\exp\left(-N^{1/2}\right) \leq \gamma_2. \end{aligned}$$

From which we deduce using (40) that for all $K \geq K(\gamma_1, \gamma_2)$ and all $N \geq N(\gamma_1, \gamma_2)$,

$$1 - \Pr\left(\mathrm{KL}\left(\pi, p^{K,N}\right) \leq \gamma_1\right) \leq \gamma_2,$$

that is

$$\lim_{K \to \infty, N \to \infty} \Pr\left(\mathrm{KL}\left(\pi, p^{K,N}\right) \leq \gamma_1\right) = 1,$$

which achieves the desired result that $\mathrm{KL}\left(\pi, p^{K,N}\right) \to 0$, in probability as $K \to \infty, N \to \infty$.

S3 .3.2 Proof of the measurability of $\mathbf{z}_{0,\mathbf{y}}^{K,N}$ (Lemma 2)

We wish to make use of the result from (Aliprantis and Border, 2006, Theorem 18.19 Measurable Maximum Theorem) to prove that we can choose a measurable function $\mathbf{y} \mapsto \mathbf{z}_{0,\mathbf{y}}^{K,N}$. More specifically this is guarantied by the following Lemma 2 which is proved below.

Background. The required materials for this lemma and the proof arise from Aliprantis and Border (2006), Chapter 18. The main concepts are recalled below.

Let f be a function on a product space $\mathcal{Y} \times \mathcal{Z}$, such that $f : \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}$. Assume that $(\mathcal{Y}, \mathcal{F})$ is a measurable space.

The function $f(\mathbf{y}, \mathbf{z})$ is said to be Caratheodory, if f is continuous in $\mathbf{z} \in \mathcal{Z}$ and measurable in $\mathbf{y} \in \mathcal{Y}$.

By definition, a correspondence ζ from a set $\mathcal{Y}$ to a set $\mathcal{Z}$ assigns each $\mathbf{y} \in \mathcal{Y}$ to a subset $\zeta(\mathbf{y}) \in \mathcal{Z}$. We write this relationship as $\zeta : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$.

A correspondence $\zeta : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$ is measurable (weakly measurable) if $\zeta^{\ell}(F) \in \mathcal{F}$ for each closed (open) subset F of $\mathcal{Z}$, where ζ^{ℓ} is the so-called lower inverse of ζ defined as $\zeta^{\ell}(F) = \{\mathbf{z} \in \mathcal{Z} : \zeta(\mathbf{y}) \cap F \neq \emptyset\}$.

Lemma 18.7 from Aliprantis and Border (2006) states the following: Suppose that $f : \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}$ is Caratheodory, where $(\mathcal{Y}, \mathcal{F})$ is a measurable space, $\mathcal{Z}$ is a metrizable space, and $\mathcal{X}$ is a topological space. For each subset H of $\mathcal{X}$, define the correspondence $\zeta_H : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$ by

$$\zeta_H(\mathbf{y}) = \{\mathbf{z} \in \mathcal{Z} : f(\mathbf{y}, \mathbf{z}) \in H\}.$$

If H is open, then ζ_H is a measurable correspondence.

Corollary 18.8 from Aliprantis and Border (2006) states the following: Suppose that $f : \mathcal{Y} \times \mathcal{Z} \to \mathcal{X}$ is Caratheodory, where $(\mathcal{Y}, \mathcal{F})$ is a measurable space, $\mathcal{Z}$ is a metrizable space, and $\mathcal{X}$ is a topological space. Define the correspondence $\zeta : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$ by

$$\zeta(\mathbf{y}) = \{\mathbf{z} \in \mathcal{Z} : f(\mathbf{y}, \mathbf{z}) = 0\}.$$

If $\mathcal{Z}$ is compact, then ζ is a measurable correspondence.

Furthermore, we have the fact that the countable unions of measurable correspondences are also measurable. We say that $\zeta : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$ admits a measurable selector, if there exists a measurable function $f : \mathcal{Y} \to \mathcal{Z}$, such that $f(\mathbf{y}) \in \zeta(\mathbf{y})$, for each $\mathbf{y} \in \mathcal{Y}$.

Theorem 18.19 (Measurable Maximum Theorem) from Aliprantis and Border (2006) then states the following. Let $\mathcal{Z}$ be a separable metrizable space and $(\mathcal{Y}, \mathcal{F})$ be a measurable space. Let $\zeta : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$ be a weakly measurable correspondence with nonempty compact values, and suppose that $f : \mathcal{Y} \times \mathcal{Z} \to \mathbb{R}$ is Caratheodory. Define $m : \mathcal{Y} \to \mathbb{R}$ by

$$m(\mathbf{y}) = \max_{\mathbf{z} \in \zeta(\mathbf{y})} f(\mathbf{y}, \mathbf{z}),$$

and define $\mu : \mathcal{Y} \twoheadrightarrow \mathcal{Z}$ to be its maximizers:

$$\mu(\mathbf{y}) = \{\mathbf{z} \in \zeta(\mathbf{y}) : f(\mathbf{y}, \mathbf{z}) = m(\mathbf{y})\}.$$

Then 1) the value function m is measurable, 2) the argmax correspondence μ has nonempty and compact values, 3) the argmax correspondence μ is measurable and admits a measurable selector.

In our context, the use of Theorem 18.19 above takes the form of Lemma 2.

Lemma 2 *Under the assumptions in Theorem 2 and with the following definitions,*

$$A^{K,N}_{\epsilon,\mathbf{y}} = \left\{\mathbf{z} \in \mathcal{Y} : D\left(p^{K,N}(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{z})\right) \leq \epsilon\right\} \text{ and } \quad A^{K,N}_{0,\mathbf{y}} = \bigcap_{\epsilon \in \mathbb{Q}_+} A^{K,N}_{\epsilon,\mathbf{y}},$$

$$B^{K,N}_{\epsilon,\mathbf{y}} = \arg\max_{\mathbf{z} \in A^{K,N}_{\epsilon,\mathbf{y}}} D_1(\pi(\cdot \mid \mathbf{z}), \pi(\cdot \mid \mathbf{y})) \text{ and } \quad B^{K,N}_{0,\mathbf{y}} = \bigcap_{\epsilon \in \mathbb{Q}_+} B^{K,N}_{\epsilon,\mathbf{y}},$$

so that $A^{K,N}_{0,\mathbf{y}} = \left\{\mathbf{z} \in \mathcal{Y} : p^{K,N}(\cdot \mid \mathbf{y}) - p^{K,N}(\cdot \mid \mathbf{z}) = 0\right\}$ *and* $B^{K,N}_{0,\mathbf{y}} = \arg\max_{\mathbf{z} \in A^{K,N}_{0,\mathbf{y}}} D_1(\pi(\cdot \mid \mathbf{z}), \pi(\cdot \mid \mathbf{y}))$.
Then, we can always choose an argmax correspondence $\mathbf{y} \twoheadrightarrow B^{K,N}_{0,\mathbf{y}}$*, which is measurable and admits a measurable selector.*

Proof of Lemma 2. Let us define the correspondence $\zeta^{K,N}_0 : \mathcal{Y} \twoheadrightarrow \mathcal{Y}$ so that $\zeta^{K,N}_0(\mathbf{y}) = A^{K,N}_{0,\mathbf{y}}$. We claim that this correspondence is a weakly measurable correspondence with nonempty compact values. Indeed, we firstly define the function $f^{K,N}(\mathbf{y}, \mathbf{z}) = p^{K,N}(\cdot \mid \mathbf{y}) - p^{K,N}(\cdot \mid \mathbf{z})$, and notice that

$$f^{K,N} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

is Caratheodory, since it is a continuous function in $\mathbf{z}$ and measurable in $\mathbf{y}$ by the continuity of $p^{K,N}$. Then, by using the (Aliprantis and Border, 2006, Corollary 18.8) and the fact that $\mathcal{Y}$ is compact, it follows that

$$\zeta^{K,N}_0(\mathbf{y}) = \left\{\mathbf{z} \in \mathcal{Y} : f^{K,N}(\mathbf{y}, \mathbf{z}) = 0\right\}$$

is measurable. Then, it is also weakly measurable (see Aliprantis and Border 2006, Lemma 18.2). Furthermore, $\zeta^{K,N}_0$ has nonempty compact values since for any $\mathbf{y} \in \mathcal{Y}$, $\zeta^{K,N}_0(\mathbf{y})$ always contains $\mathbf{y}$, and $\zeta^{K,N}_0(\mathbf{y}) = \left[f^{K,N}(\mathbf{y}, \cdot)\right]^{-1}(\{0\})$ is a compact set since the inverse image of continuous function $f^{K,N}(\mathbf{y}, \cdot)$ of compact set is also compact.

Then, since we assume that $(\mathbf{y}, \mathbf{z}) \mapsto D_1(\pi(\cdot \mid \mathbf{z}), \pi(\cdot \mid \mathbf{y}))$ is a continuous function in $\mathbf{z}$ and measurable in $\mathbf{y}$, then it is also a Caratheodory function. We also remark that $B^{K,N}_{0,\mathbf{y}}$ can be written as a argmax correspondence

$$B^{K,N}_{0,\mathbf{y}} = \arg\max_{\mathbf{z} \in \zeta^{K,N}_0(\mathbf{y})} D_1(\pi(\cdot \mid \mathbf{z}), \pi(\cdot \mid \mathbf{y})).$$

By using the result from Aliprantis and Border, 2006, Theorem 18.19, Measurable Maximum Theorem, we conclude that the the argmax correspondence $B_{0,\mathbf{y}}^{K,N}$ is measurable and admits a measurable selector, that is, we can always choose a measurable function $\mathbf{y} \mapsto \mathbf{z}_{0,\mathbf{y}}^{K,N} \in B_{0,\mathbf{y}}^{K,N}$.

S4 Additional illustrations

S4 .1 Bivariate Beta model

The bivariate beta model proposed by Crackel and Flegal (2017) is defined with five positive parameters $\theta_1, \ldots, \theta_5$ by letting

$$v_1 = \frac{u_1 + u_3}{u_5 + u_4}, \text{ and } v_2 = \frac{u_2 + u_4}{u_5 + u_3}, \tag{41}$$

where $u_i \sim \text{Gamma}(\theta_i, 1)$, for $i \in [5]$, and setting $z_1 = v_1/(1 + v_1)$ and $z_2 = v_2/(1 + v_2)$. The bivariate random variable $\mathbf{z}^\top = (z_1, z_2)$ has marginal laws $z_1 \sim \text{Beta}(\theta_1 + \theta_3, \theta_5 + \theta_4)$ and $z_2 \sim \text{Beta}(\theta_2 + \theta_4, \theta_5 + \theta_3)$. We perform ABC using samples of size $R = 100$. The observed sample is generated from the model with true parameter values $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (1, 1, 1, 1, 1)$. The prior on each of the model parameters is taken to be independent and uniform over interval $[0, 5]$.

Figure S1 shows the marginal ABC posterior distributions for each of the 5 parameters and comparing 5 ABC procedures.

We then summarise each sample using 14 quantiles and apply the ABC procedures on these summarised data sets. The marginal posteriors are shown in Figure S2 for the 5 parameters and the 5 procedures. The GLLiM-MW2-ABC procedure based on the MW_2 distance is the best while the one based on L_2 performs very poorly. This is surprising as both methods are based on the same GLLiM surrogates. Understanding the failure of the L_2 distance in this specific case would require more investigations. The GLLiM mixture for the observation to be inverted is also shown and exhibits modes near the true parameters values. GLLiM-MW2-ABC and GLLiM-E-ABC perform similarly while the addition of log-variances in GLLiM-EV-ABC does not seem to improve posterior estimation. GLLiM-E-ABC and the semi-automatic method both rely on an estimation of the posterior means but show posteriors of different shapes in particular for θ_1 and θ_4 (Figure S2). Differences between the two methods are also observed in Figure S1 for θ_2 and θ_4 but in this case the two methods do not used the same summaries.

For a more quantitative comparison, we compute for each posterior samples of size S, empirical means of the parameters, $\bar{\theta}_j = \frac{1}{S}\sum_{i=1}^{S} \theta_j^i$, and empirical root mean square errors (RMSE) defined as $R(\theta_j) = \sqrt{\frac{1}{S}\sum_{i=1}^{S}(\theta_j^i - \theta_j^0)^2}$ where $j \in [5]$, $S = 50$ and θ_j^i is the sample i for θ_j and θ_j^0 is the true parameter value. Table S1 shows these quantities for the posteriors shown in Figures S1 and S2.

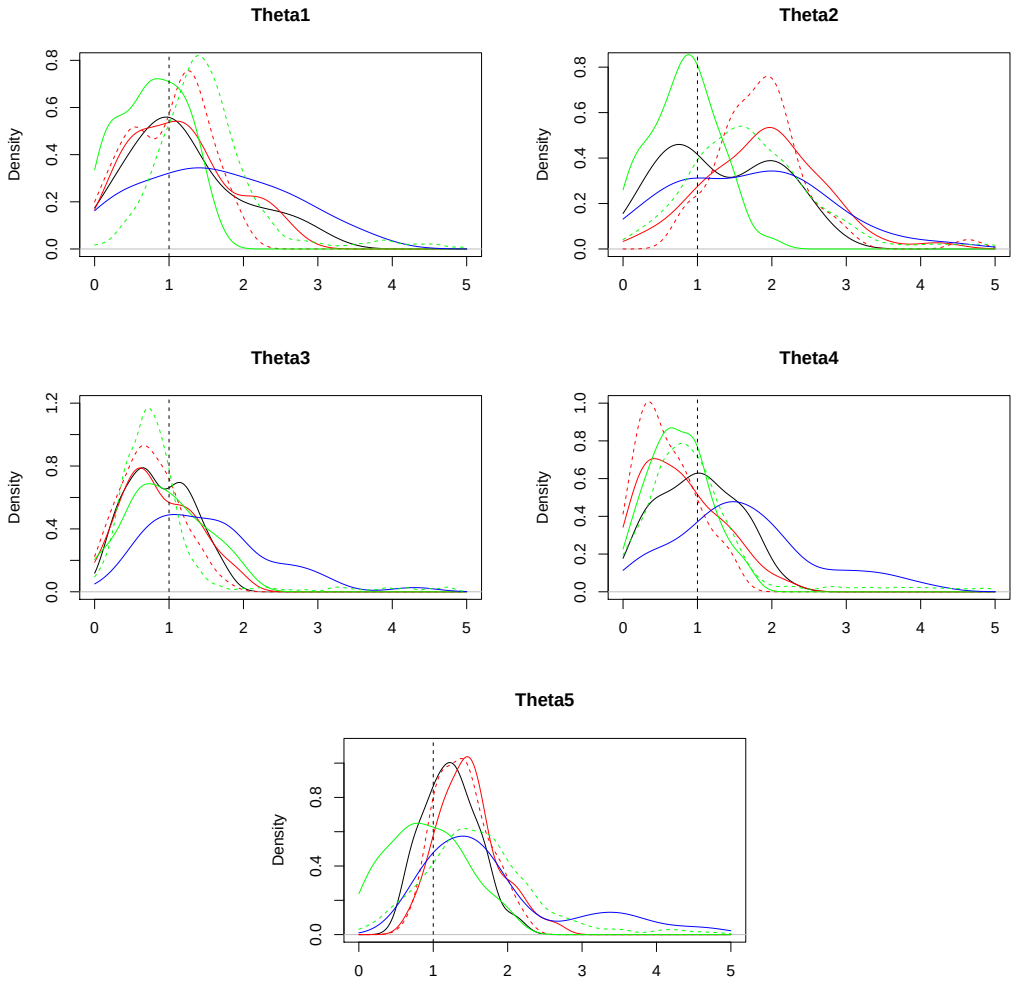


Figure S1: Bivariate Beta model: marginal posteriors Marginal for parameters $\theta_1, \ldots, \theta_5$. Realisations are made of $R = 100$ *i.i.d.* observations. ABC procedures are applied on these observations, GLLiM-E-ABC (red), GLLiM-EV-ABC (dotted red), GLLiM-MW2-ABC (black), GLLiM-L2-ABC (blue), except for semi-automatic ABC which uses reduced observations to 14 quantiles (green). The GLLiM mixture is also shown (dotted green). GLLiM-iid is applied with $K = 100$. The black dashed lines indicate the true parameter values.

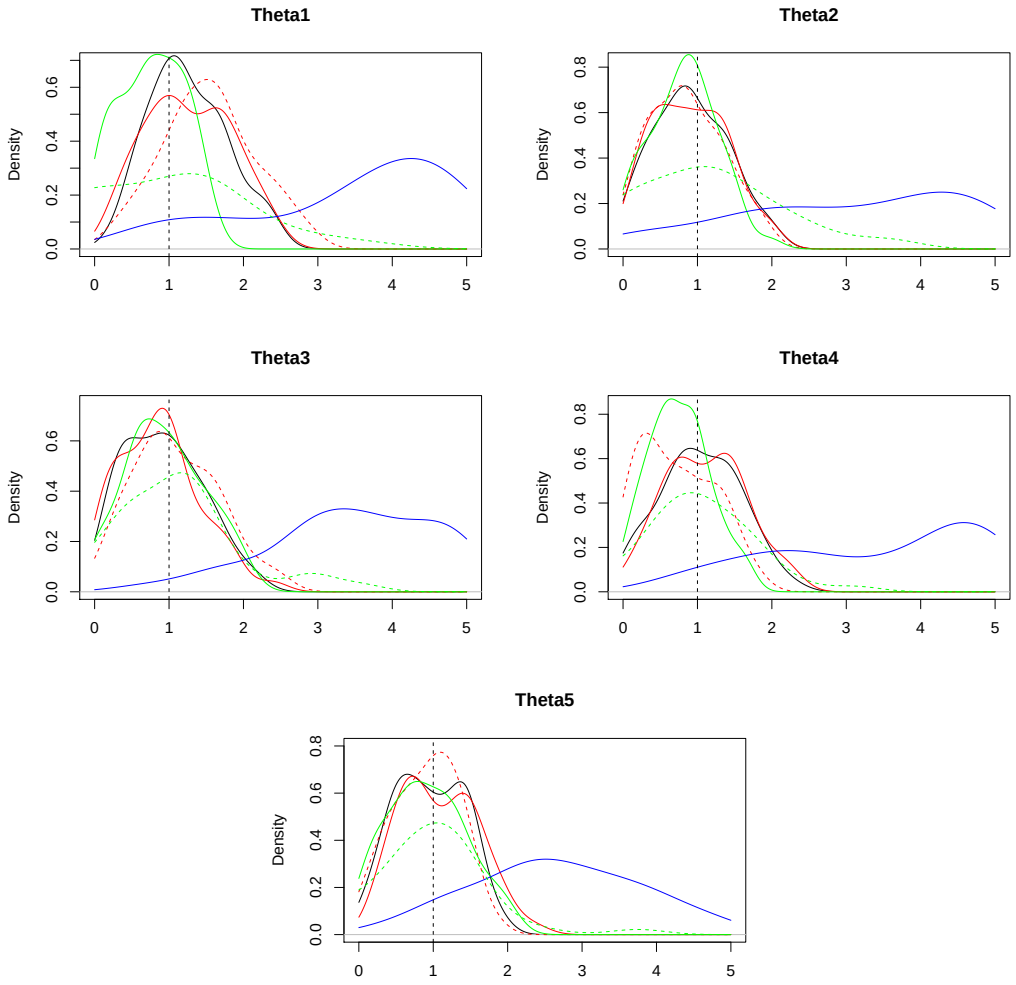


Figure S2: Bivariate Beta model: marginal posteriors for parameters $\theta_1, \ldots, \theta_5$. Each set of $R = 100$ *i.i.d.* realisations has been reduced to 14 quantiles. All ABC procedures are applied on these reduced observations: semi-automatic ABC (green), GLLiM-E-ABC (red), GLLiM-EV-ABC (dotted red), GLLiM-MW2-ABC (black), GLLiM-L2-ABC (blue), semi-automatic ABC (green). The corresponding GLLiM mixture is also shown (dotted green). Standard GLLiM is applied with $K = 40$. The black dashed lines indicate the true parameter values.

S4 .2 Moving average model

The MA(2) process is a stochastic process $(y'_t)_{t\in\mathbb{N}^*}$ defined by

$$y'_t = z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2}, \tag{42}$$

Procedure	$\bar{\theta}_1$	$\bar{\theta}_2$	$\bar{\theta}_3$	$\bar{\theta}_4$	$\bar{\theta}_5$	R(θ_1)	R(θ_2)	R(θ_3)	R(θ_4)	R(θ_5)
GLLiM mixture	1.504	1.736	**0.890**	0.989	1.616	0.926	1.276	0.824	0.848	1.021
GLLiM-E-ABC	**1.142**	1.899	0.871	0.786	1.472	0.678	1.181	0.498	0.568	0.626
GLLiM-EV-ABC	0.990	1.867	0.746	0.594	1.385	**0.505**	1.077	0.469	0.562	0.517
GLLiM-L2-ABC	1.597	1.700	1.534	1.627	1.827	1.1445	1.224	0.968	1.117	1.295
GLLiM-MW2-ABC	1.211	**1.319**	0.872	**1.004**	**1.235**	0.790	**0.820**	**0.439**	**0.523**	**0.426**
	with 14 quantiles as summaries									
Semi-auto ABC	**0.770**	0.825	**0.947**	0.756	0.917	**0.493**	**0.472**	**0.524**	**0.468**	0.523
GLLiM mixture	0.448	0.858	0.739	0.552	0.577	1.685	1.464	1.367	1.300	1.214
GLLiM-E-ABC	1.266	0.905	0.872	1.105	1.082	0.629	0.501	0.550	0.541	0.514
GLLiM-EV-ABC	1.530	0.852	1.095	0.727	0.904	0.808	0.504	0.577	0.565	**0.450**
GLLiM-L2-ABC	3.390	3.010	3.467	3.361	2.653	2.747	2.492	2.693	2.732	1.995
GLLiM-MW2-ABC	1.257	**0.909**	0.921	**1.042**	**0.950**	0.573	0.500	**0.524**	0.529	0.464

Table S1: Bivariate Beta model: parameter means, and RMSE ($R(\cdot)$) for ABC posterior distributions shown in Figures S1 and S2 when the observed data is generated with $\boldsymbol{\theta} = (1, 1, 1, 1, 1)$. The ABC posterior values are computed as empirical values over samples of size 50. Means closest to 1 and best (lowest) RMSE values are in boldface.

where $\{z_t\}$ is an *i.i.d.* sequence, according to a standard normal distribution and θ_1 and θ_2 are scalar parameters. A standard identifiability condition is imposed on this model leading to a prior distribution on the triangle described by the inequalities

$$-2 < \theta_1 < 2, \quad \theta_1 + \theta_2 > -1, \quad \theta_1 - \theta_2 < 1 \ .$$

As in most papers, the prior on the two model parameters is taken uniform over the triangular domain. Natural summary statistics for this model are the empirical auto-covariances of lag 1 and 2, which converge to a one-to-one function of the two parameters. This example is a way to illustrate our method on time series in the same manner as Bernton et al. (2019). Their Wasserstein-ABC proposal uses empirical distributions and, like other data discrepancy based methods, is in principle only valid for *i.i.d.* observations. However, they also investigate the use of the method to time series where observations are not *i.i.d.*. We make a similar attempt in this work and show how it can be interpreted in our framework.

For each pair of parameters (θ_1, θ_2) in the triangular domain, a series of length 150 is simulated according to the MA(2) model (42). We consider time series of length 150, instead of 100 in Jiang et al. (2017). This is repeated N times so that the number of pairs in the learning set is $N = 10^5$. The series to be inverted is simulated similarly with true parameters $\theta_1 = 0.6$ and $\theta_2 = 0.2$. To learn a GLLiM model with $d = 150$, $\ell = 2$, $K = 30$, and no constraints on the covariance matrices for the likelihood part of the model, requires the estimation of 353429 parameters. To reduce the model complexity while going beyond the alternative isotropic or diagonal cases, we propose to use the *i.i.d.* adaptation of GLLiM.

Table S2: MA(2) model: parameter means, standard deviations and correlations for the exact and ABC posterior distributions shown in Figure S3 when the observed data is generated with $\theta_1 = 0.6$ and $\theta_2 = 0.2$. The exact posterior values are computed numerically, while the ABC posterior values are computed as empirical values over samples of size 100. Closest values to the true posterior ones are in bold.

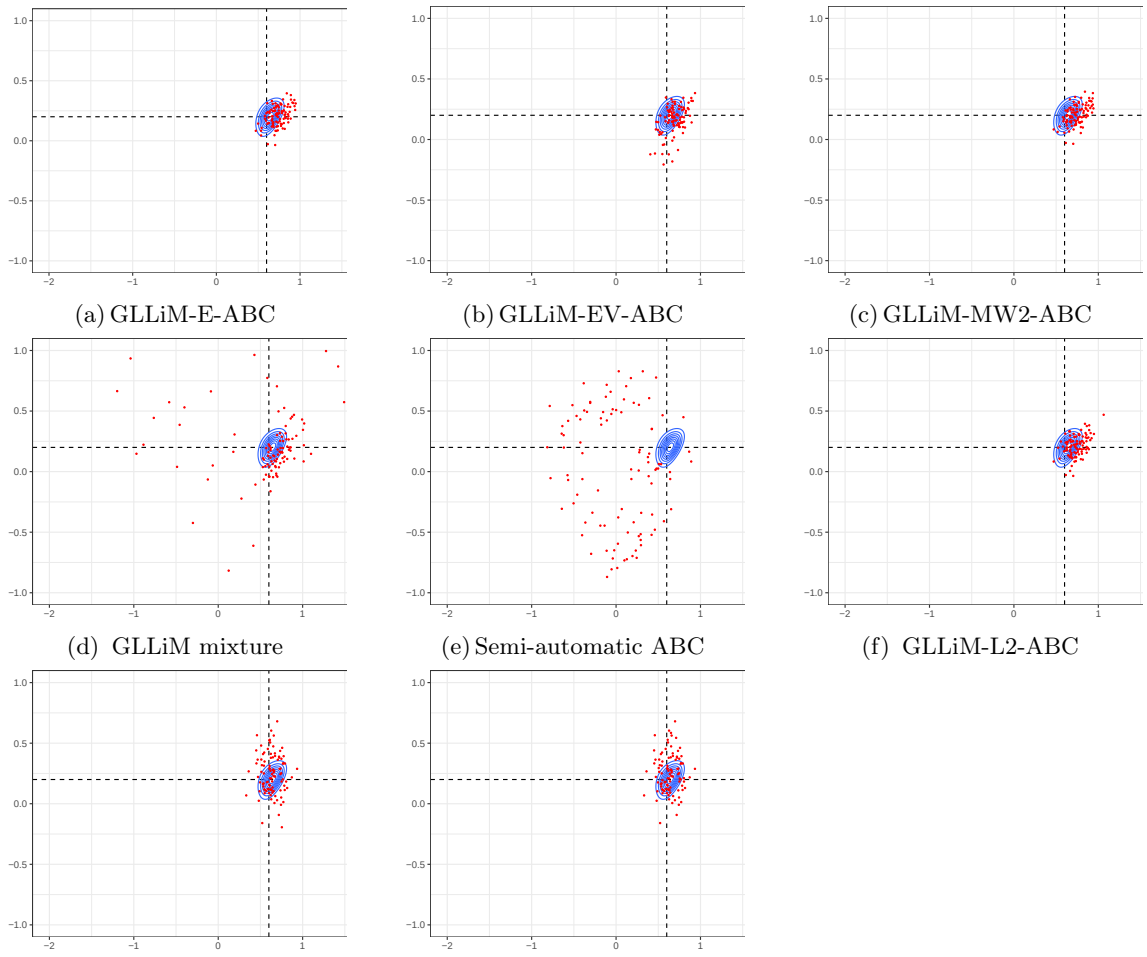
Posterior	mean(θ_1)	mean(θ_2)	std(θ_1)	std(θ_2)	cor(θ_1, θ_2)
Exact	0.635	0.203	0.080	0.076	0.472
Auto-cov Rejection ABC	**0.635**	0.246	0.107	0.164	0.018
Auto-cov Semi-auto	0.637	0.250	0.109	0.159	-0.045
Semi-auto ABC	0.026	0.027	0.406	0.476	-0.112
GLLiM mixture	0.521	0.182	0.499	0.294	-0.007
GLLiM-E-ABC	0.737	0.208	0.104	0.084	0.454
GLLiM-EV-ABC	0.689	0.163	0.110	0.120	**0.474**
GLLiM-L2-ABC	0.740	0.213	**0.103**	0.088	0.436
GLLiM-MW2-ABC	0.742	**0.206**	0.104	**0.084**	0.527

GLLiM is applied with $d = 3$, $R = 50$ and no constraint on the 3×3 blocs themselves (629 parameters). A second experiment is made with $R = 5$ and $d = 30$ *i.e.* with 5 blocs of size 30×30 with no constraint on the bloc structure (16829 parameters) . The second setting is retained as it shows better precision on θ_2 in particular. In terms of approximation this is equivalent to neglect only few correlations in the GLLiM approximation of the likelihood.

For ABC procedures, the tolerance threshold ϵ is set to the 0.1% quantile leading to selected samples of size 100. Empirical values for parameter means, standard deviations and correlation when applying the different ABC schemes for one observed time series are compared to the true ones computed numerically. The true posterior means, standard deviations of θ_1 and θ_2 are computed numerically using importance sampling. The true posterior correlation between θ_1 and θ_2 is also computed this way. The corresponding ABC estimations and samples are shown in Table S2 and Figure S3. The results are qualitatively similar to that of Jiang et al. (2017) with a poor estimation of the means for the semi-automatic ABC procedure on the full time series. They also confirm results already observed in previous works, namely that semi-automatic and auto-covariance-based procedures do not well capture correlation information between θ_1 and θ_2. Surprisingly, the GLLiM mixture approximation of the posterior also provides a poor estimation of the correlation but this estimation improves a lot when adding the ABC step.

S4 .3 Non-identifiable models

Our main targets are posterior distributions with multiple modes for which our method is more likely to provide significantly better performance than existing approaches. It is straightforward to construct models that lead to multimodal posteriors by considering



(a) GLLiM-E-ABC (b) GLLiM-EV-ABC (c) GLLiM-MW2-ABC

(d) GLLiM mixture (e) Semi-automatic ABC (f) GLLiM-L2-ABC

(g) Auto-covariances (h) Semi-automatic auto-covariances

Figure S3: MA(2) model. A GLLiM-iid model is learned on a data set of size $N = 10^5$ with $K = 30$, $d = 30$, $R = 5$. Selected samples (100 points) using (a) GLLiM posterior expectations, (b) GLLiM posterior expectations and log variances, (c) MW_2 distances, (d) the approximate GLLiM posterior for the observed data, (e) semi-automatic ABC, (f) L_2 distances, (g) Auto-covariances as summary statistics, (h) semi-automatic ABC on auto-covariances, Contours of the true posterior distribution computed numerically are shown in blue. The true parameters are 0.6 and 0.2 as indicated by the dashed lines.

likelihoods that are invariant by some transformation.

S4 .3.1 Ill-posed inverse problems

Here, we consider inverse problems for which the solution is not unique. This setting is quite common in practice and can occur easily when the forward model exhibits some invariance, *e.g.*, when considering the negative of the parameters. A simple way to model this situation consists of assuming that the observation $\mathbf{y}$ is generated as a realization of

$$\mathbf{y} = F(\boldsymbol{\theta}) + \boldsymbol{\varepsilon},$$

where F is a deterministic theoretical model coming from experts and $\boldsymbol{\varepsilon}$ is a random variable expressing the uncertainty both on the theoretical model and on the measurement process. A common assumption is that $\boldsymbol{\varepsilon}$ is distributed as a centered Gaussian noise. Non-identifiability may then arise when $F(-\boldsymbol{\theta}) = F(\boldsymbol{\theta})$. Following this generative approach, a first simple example is constructed with a Student t-distributed noise leading to the likelihood:

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

where $\mathcal{S}_d(\cdot; \mu^2 \mathbb{1}_d, \sigma^2 \mathrm{I}_d, \nu)$ is the pdf of a d-variate Student t-distribution with a d-dimensional location parameter with all dimensions equal to μ^2, diagonal isotropic scale matrix $\sigma^2 \mathrm{I}_d$ and degree-of-freedom (dof) parameter ν. Recall that for a Student t-distribution, a diagonal scale matrix is not inducing independent dimensions so that $\mathbf{y}$ is not a set of *i.i.d.* univariate Student t observations. The dof controls the tail heaviness; *i.e.*, the smaller the value of ν, the heavier the tail. In particular, for $\nu \leq 2$, the variance is undefined, while for $\nu \leq 1$ the expectation is also undefined. In this example, we set $\sigma^2 = 2$, $\nu = 2.1$, and μ is the parameter to estimate.

For all compared procedures, we set $d = 10$, $K = 10$, $N = M = 10^5$, and the tolerance level ϵ to the 0.1% quantile of observed distances, so that all selected posterior samples are of size 100.

Figure S4 shows the true and the compared ABC posterior distributions for a 10-dimensional observation $\mathbf{y}$, simulated under a process with $\mu = 1$. The true posterior exhibits the expected symmetry with modes close to the values: $\mu = 1$ and $\mu = -1$. The simple rejection ABC procedure based on GLLiM expectations (GLLiM-E-ABC) and the semi-automatic ABC procedure both show over dispersed samples with wrongly located modes. The GLLiM-EV-ABC exhibits two well located modes but does not preserve the symmetry of the true posterior. The distance-based approaches, GLLiM-L2-ABC and GLLiM-MW2-ABC both capture the bimodality. GLLiM-MW2-ABC is the only method to estimate a symmetric posterior distribution with two modes of equal importance. Note, however, that in term of precision, the posterior distribution estimation remains difficult considering an observation of size only $d = 10$.

This simple example shows that the expectation as a summary statistic suffers from the presence of two equivalent modes, while the approaches based on distances are more

robust. There is a clear improvement in complementing the summary statistics with the log-variances. Although in this case, this augmentation provides a satisfying bimodal posterior estimate, it lacks the expected symmetry of the two modes. The GLLiM-MW2-ABC procedure has the advantage of exhibiting a symmetric posterior estimate, that is more consistent with the true posterior.

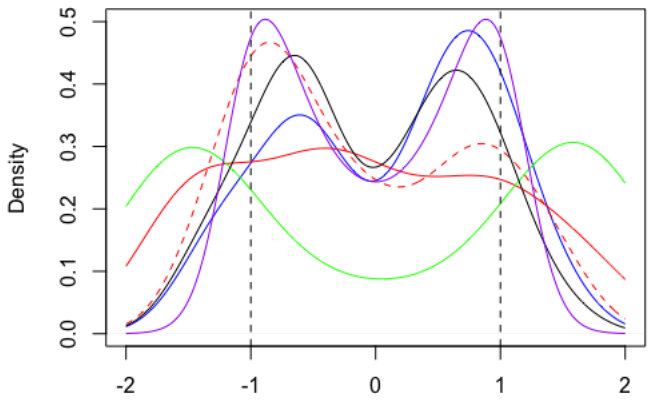


Figure S4: Non identifiable Student t-distribution. ABC posterior distributions from the selected samples. GLLiM-L2-ABC in blue, GLLiM-MW2-ABC in black, semi-automatic ABC in green, GLLiM-E-ABC (expectations) in red and GLLiM-EV-ABC (expectations and log-variances) in dotted red line. The true posterior is shown in purple. The dashed lines indicate the μ (equivalent) values used to generate the observation.

In the following subsection we present another case that cannot be cast as the above generating process but also exhibit a transformation invariant likelihood.

S4 .3.2 Sum of moving average models of order 2 (MA(2))

Using the same MA(2) process as already defined in Section S4 .2, we consider a transformation that consists of taking the opposite sign of θ_1 and keeping θ_2 unchanged. The considered observation corresponds then to a series obtained by summing the two MA models, defined below

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

where $\{z_t\}$ and $\{z'_t\}$ are both *i.i.d.* sequences, generated from a standard normal distribution. It follows that a vector of length d, $\mathbf{y} = (y_1, \ldots, y_d)^\top$, is distributed according to a multivariate d-dimensional centered Gaussian distribution with a Toeplitz covariance matrix whose first row is $(2(\theta_1^2 + \theta_2^2 + 1), 0, 2\theta_2, 0, \ldots, 0)$. The likelihood is therefore invariant by the transformation proposed above, and so is the uniform prior over the triangle.

It follows that the posterior is also invariant by the same transformation and can then be chosen so as to exhibit two symmetric modes.

For all procedures, we set $K = 80$ and $N = M = 10^5$, and ϵ to the 1% distance quantile, so that all selected posterior samples are of size 1000. An observation of size $d = 10$ is simulated from the model with $\theta_1 = 0.7$ and $\theta_2 = 0.5$. ABC posterior distribution estimates are shown in Figure S5.

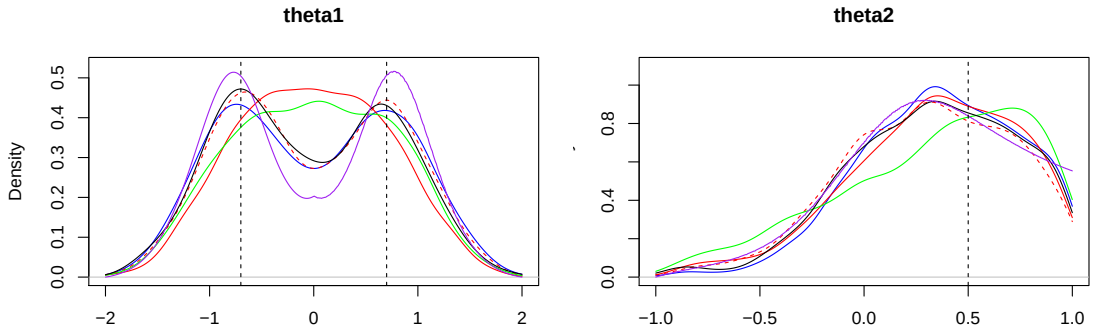


Figure S5: Sum of MA(2) models. Posterior marginals from the samples selected with a 1% quantile (1000 values): semi-automatic ABC (green), GLLiM-L2-ABC (blue), GLLiM-MW2-ABC (black), GLLiM-E-ABC (red) and GLLiM-EV-ABC (dotted red). The true marginal posteriors are shown in purple. The dashed lines show the values used to simulate the observation $\theta_1 = 0.7$ and $\theta_2 = 0.5$.

The level sets of the true posterior can be computed from the exact likelihood and a grid of values for θ_1 and θ_2. For the setting used here, none of the considered ABC procedures is fully satisfactory, in that the selected samples are all quite dispersed. This is mainly due to the relatively low size of the observation ($d = 10$). The tests made in Section S4 .2 with a much larger $d = 150$ provided in contrast very satisfying samples as visible in supplementary Figure S3. This can also be observed in Marin et al. (2012) (Figures 1 and 2), where ABC samples are less dispersed for a size of $d = 100$ and quite spread off when d is reduced to $d = 50$, even when the autocovariance is used as summary statistic.

Despite the relative spread of the parameters accepted after the ABC rejection, the posterior marginals, shown in Figure S5, provide an interesting comparison. GLLiM-D-ABC and GLLiM-EV-ABC procedures show symmetric θ_1 values, in accordance with the symmetry and bimodality of the true posterior. The use of the L_2 or MW_2 distances does not lead to significant differences. GLLiM-E-ABC and semi-automatic ABC behave similarly and do not capture the bimodality on θ_1, but the addition of the posterior log-variances in GLLiM-EV-ABC improves on GLLiM-E-ABC. These results suggest that although GLLiM may not provide good approximations of the first posterior moments, it can still provide good enough approximations of the surrogate posteriors in GLLiM-D-ABC. For θ_2, all

posteriors are rather close to the true posterior marginal except for semi-automatic ABC which shows a mode at a wrong location when compared to the true posterior.

A similar example using MA(1) processes is also provided in the next subsection.

S4 .3.3 Sum of moving average models of order 1 (MA(1))

The MA(1) process is a stochastic process $(y'_t)_{t\in\mathbb{N}^*}$ defined by

$$y'_t = z_t + \rho z_{t-1} \ .$$

In order to construct bimodal posterior distributions, we consider the following sum of two such models. At each discrete time step t we define,

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

where $\{z_t\}$ and $\{z'_t\}$ are both *i.i.d.* sequences, according to a standard normal distribution and ρ is an unknown scalar parameter. It follows that a vector of length d, $\mathbf{y} = (y_1, \ldots, y_d)^\top$ is distributed according to a multivariate d-dimensional centered Gaussian distribution with an isotropic covariance matrix whose diagonal entries are all equal to $2(\rho^2 + 1)$. The likelihood is therefore invariant by symmetry about 0 and so is the prior on ρ assumed to be uniform over $[-2, 2]$. It follows that the posterior on ρ is also invariant by this transformation and can thus be chosen so as to exhibit two symmetric modes. The true posterior looks similar to the one in Section S4 .3 but ρ is now a parameter impacting upon the variance of the likelihood.

For all procedures, we set $N = M = 10^5$, and ϵ to the 0.1% quantile of observed distances so that all selected posterior samples are of size 100. In terms of difficulty, the main difference with the example in Section S4 .3 lies in a higher non-linearity of the likelihood and of the model joint distribution. We then report results with a higher choice of $K = 20$. When $K = 10$, results are similar except for GLLiM-EV-ABC, which does not show improvement over GLLiM-E-ABC.

A $d = 10$ dimensional observation, simulated from a process with $\rho = 1$, is considered. The ABC posterior distributions derived from the selected samples are shown for each of the compared procedures in Figure S6. The expectation-based summary statistics approaches (semi-automatic ABC and GLLiM-E-ABC) do not capture the bimodality. Adding the posterior log-variances (red dotted line) allows to recover the two modes. GLLiM-EV-ABC, GLLiM-MW2-ABC and GLLiM-L2-ABC provide similar bimodal posterior distributions, with more symmetry between the two modes for the two first methods.

S4 .4 Sound source localization

S4 .4.1 Two-microphone setup

Considering the two-microphone configuration described in Section 6.3.1 in the manuscript, we compare the four ABC methods using GLLiM with semi-automatic ABC. Recall that

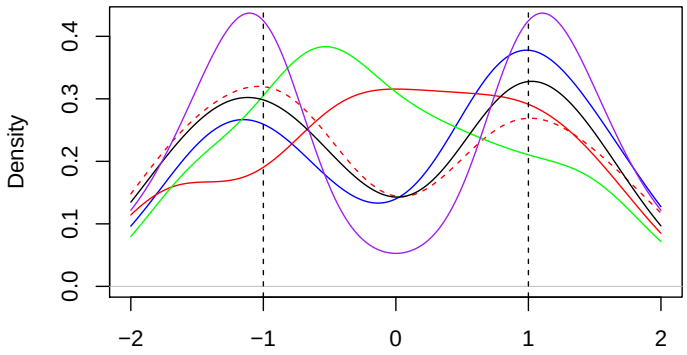


Figure S6: Sum of MA(1) models. ABC posterior distributions from the selected samples. GLLiM-L2-ABC in blue, GLLiM-MW2-ABC in black, semi-automatic ABC in green, GLLiM-E-ABC (expectations) in red and GLLiM-EV-ABC (expectations and log-variances) in dotted red line. The true posterior is shown in purple. The dashed lines indicate the ρ (equivalent) values used to generate the observation.

the likelihood is defined by

$$f_{\boldsymbol{\theta}}(\mathbf{y}) = \mathcal{S}_d(\mathbf{y}; \mathrm{ITD}(\boldsymbol{\theta})\mathbb{I}_d, \sigma^2 \mathrm{I}_d, \nu) \; . \tag{43}$$

In this example, the semi-automatic ABC procedure uses the same data set for both the regression and rejection steps. For a fair comparison, we thus also learn here a GLLiM model from the same data set. We use a training set of $M = 10^6$ pairs $(\boldsymbol{\theta}, \mathbf{y}) \in \Theta \times \mathbb{R}^{10}$, simulated from a uniform distribution on Θ and applying model (43). The estimated GLLiM model consists of $K = 20$ Gaussian components with an isotropic constraint. The ABC procedures are then run on the same $M = 10^6$ training set. A selected set of 1000 samples is retained by thresholding the distances under the 0.1% quantile.

Figure S7 shows the ABC samples with another sample simulated from the GLLiM posterior distribution, corresponding to the observation $\mathbf{y}$ (Figure S7 (d)). This GLLiM posterior is a 20-component Gaussian mixture. Another sample obtained using the Metropolis–Hastings algorithm, as implemented in the R package **mcmc** (Geyer and Jonhson, 2020), is shown in Figure S7 (g)). Figure S7 (h) show the true posterior around hyperboloids, which are symmetric with respect to the microphones line and its mediatrice, and contain the true sound source localization as expected.

All tested procedures except semi-automatic ABC reflect the bimodality of the posterior distribution. The 20-component GLLiM mixture (Figure S7 (d)) reproduces correctly the bimodality of the true posterior. However, the accuracy is improved when using an additional ABC step. GLLiM-EV-ABC, GLLiM-L2-ABC and GLLiM-MW2-ABC lead to very similar selected samples (Figure S7 (b,c,f)). Using only the GLLiM posterior expectations as summary statistics is less informative although the GLLiM mixture itself appears as a

reasonable approximation that well captures the main shape of the true posterior. Interestingly, semi-automatic ABC and GLLiM-E-ABC provide different selections, although both procedures are based on a preliminary estimation of the posterior means. In this example, the true posterior means are all zero due to symmetry in the posterior distributions. The semi-automatic ABC selected sample is then the one expected as the true posterior means do not carry any information on the parameter values. The posterior means approximated by GLLiM are also all around zero but the structure visible in the selected sample suggests that the surrogate means still capture some information on the parameter values, probably through the estimation bias. Paradoxically the poor semi-automatic ABC selection may be due to a more accurate preliminary regression step.

S4 .4.2 Two-microphone pairs

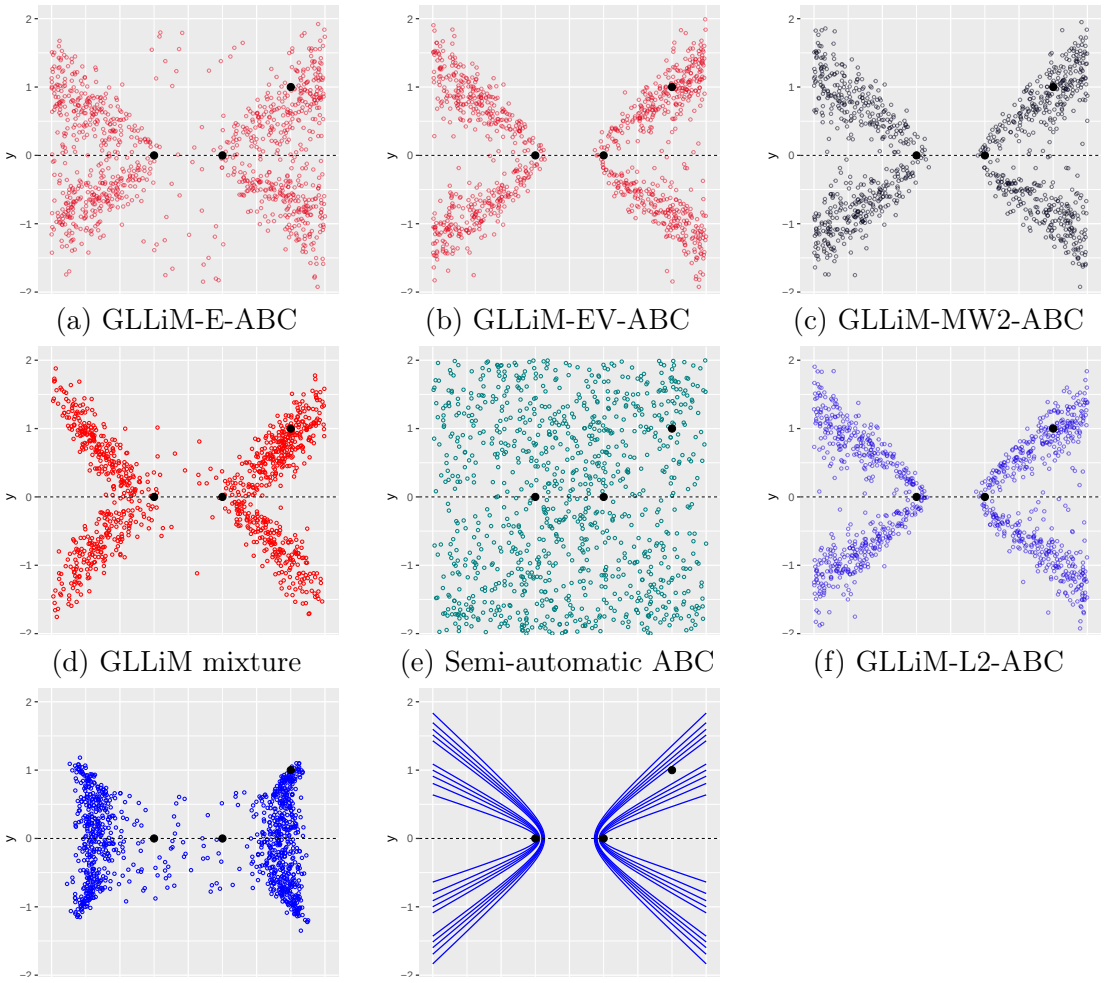
When a larger data set with $M = 10^6$ is used to learn GLLiM as it is done to fit the semi-automatic ABC regression, Figure S8 shows that both GLLiM-E-ABC and GLLiM-EV-ABC improve. A more accurate GLLiM fit may therefore have an impact on this latter procedures while GLLiM-D-ABC procedures are less sensitive to the quality of the fit.

S4 .5 Planetary science example

S4 .5.1 Synthetic data from the Hapke model

Prior to real data inversion, to illustrate the performance of the procedures, we consider an observation simulated from the Hapke model as explained above. As already mentioned the Hapke model is quite difficult to invert due to equivalent solutions and low sensitivity of the model to some of the parameters. Therefore as a first validation and for a useful comparison of the procedures we chose to invert a simulated observation as close as possible to the real observed signal described in the next section. Among the simulated signals, in the ABC data set, we chose then the one whose correlation with the real observed one was the highest. This synthetic signal has been generated from the Hapke model applied to parameter values $(\omega, \bar{\theta}, b, c) = (0.68, 0.04, 0.23, 0.04)$, with an additional Gaussian noise with standard deviation of 0.05.

Figure S9 shows the marginal posteriors obtained for each parameter using the five ABC procedures and for different tolerance values ϵ chosen as the 0.05%, 0.1% and 1% quantiles of the observed distances. A particular feature of this synthetic example is the relatively low value of $\overline{\theta}$, which does not correspond to a value expected in real data. Experts consider that reasonable values for $\overline{\theta}$ are between 0.33 and 0.66 (representing in the original space an angle between 10 and 20 degrees). The Hapke model is also such that ω and $\overline{\theta}$ values can interact to allow the reconstruction of a given spectrum. In Figure S9, this effect is visible on the slightly shifted modes of the posterior distributions for ω and $\bar{\theta}$ compared to the value used for the simulation. This bias is compensating for the overly small value of $\overline{\theta}$. Then the fact that posterior distributions for c are sharper than those for b is also



(a) GLLiM-E-ABC (b) GLLiM-EV-ABC (c) GLLiM-MW2-ABC

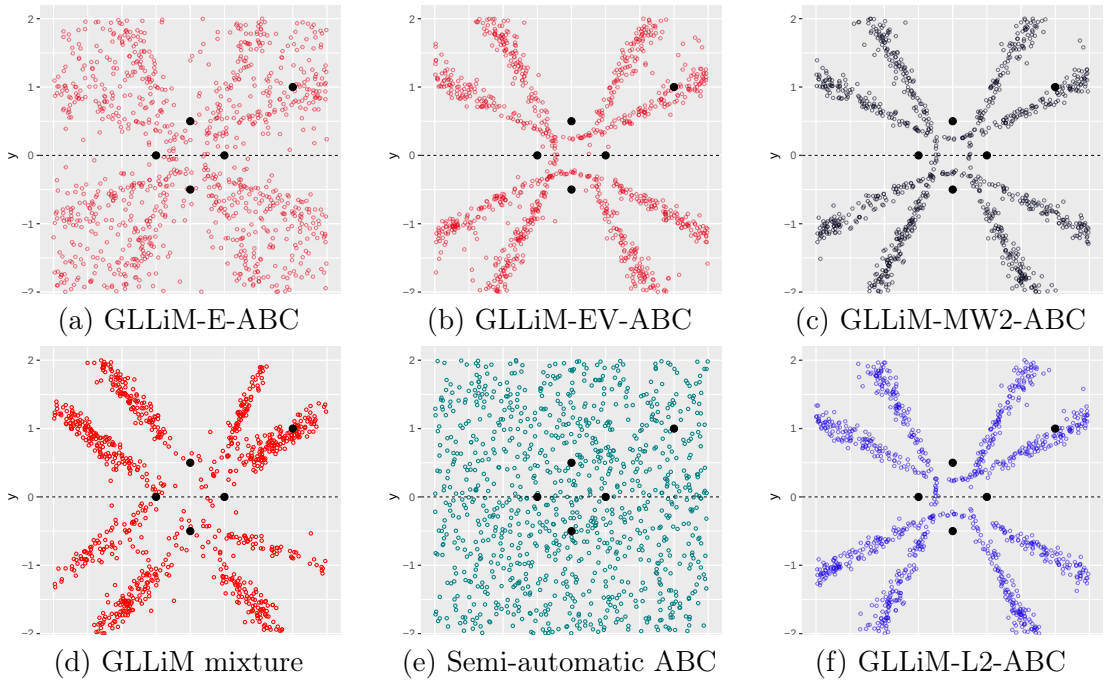
(d) GLLiM mixture (e) Semi-automatic ABC (f) GLLiM-L2-ABC

(g) Metropolis-Hastings (h) True posterior

Figure S7: Sound source localization. GLLiM and semi-automatic ABC both use the same large data set of size $M = 10^6$. Selected samples using (a) GLLiM posterior expectations, (b) GLLiM posterior expectations and log variances, (c) MW_2 distances, (d) the approximate GLLiM posterior for the observed data, (e) semi-automatic ABC, (f) L_2 distances, (g) Metropolis-Hastings sample and (h) contours of the true posterior distribution. Black points on the dotted line are the microphones positions. The third black point is the true sound source localization.

consistent with expert knowledge according to which b and $\overline{\theta}$ are more difficult to estimate than ω and c.

More generally, this example highlights the performance of the different ABC methods.



(a) GLLiM-E-ABC (b) GLLiM-EV-ABC (c) GLLiM-MW2-ABC

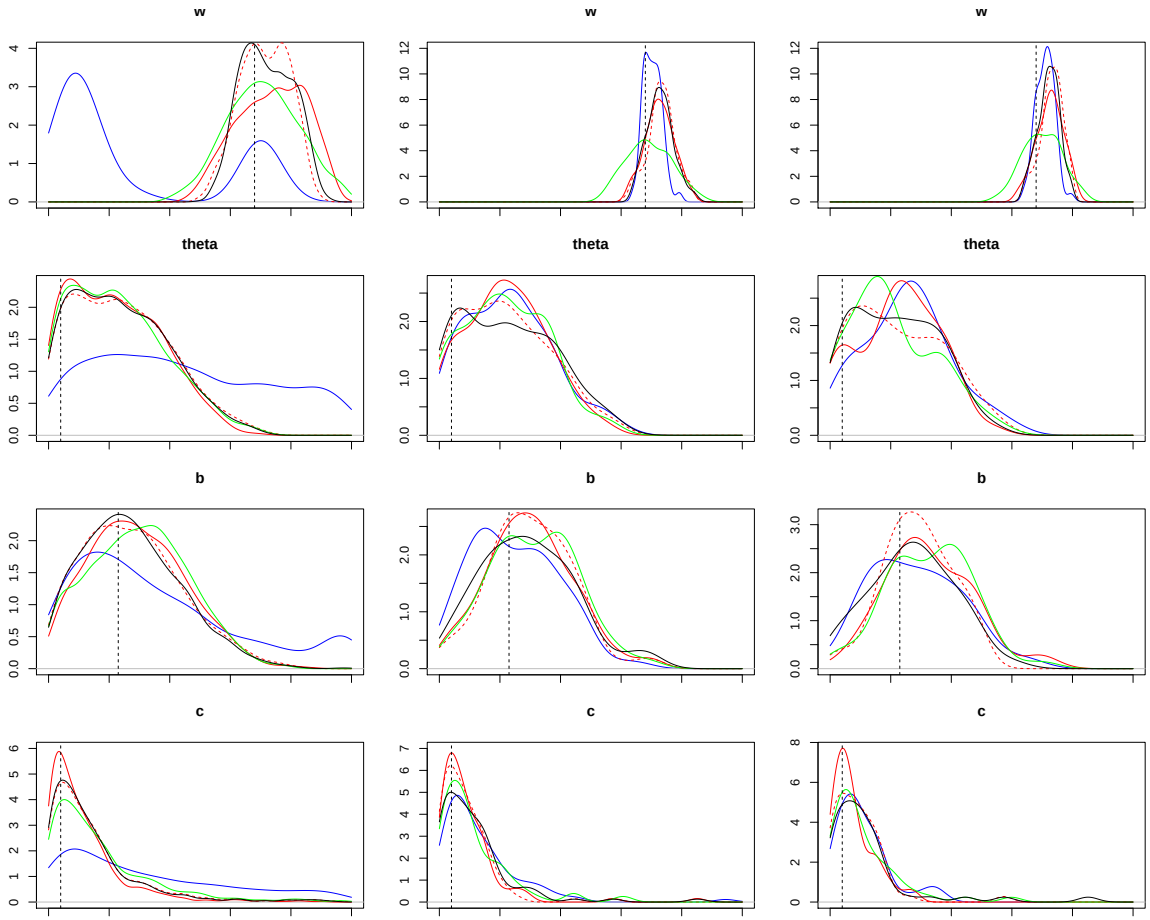
(d) GLLiM mixture (e) Semi-automatic ABC (f) GLLiM-L2-ABC

Figure S8: Sound source localization with a mixture of two microphones pairs. GLLiM is learned with the largest data set of size $M = 10^6$. Selected samples using (a) GLLiM posterior expectations, (b) GLLiM posterior expectations and log variances, (c) MW_2 distances, (d) the approximate GLLiM posterior for the observed data, (e) semi-automatic ABC, (f) L_2 distances. Black points on the dotted line are the microphones positions. The fifth black point is the true sound source localization.

It is interesting to vary ϵ to observe the behavior of the different methods. A lower ϵ can be used to check if one of the modes may vanish (*i.e.* with a more drastic thresholding) or is confirmed when the selection is more permissive. The GLLiM-L2-ABC procedure seems less robust, than the other procedures, to these variations and even degrades in performance when the thresholding is too permissive. The two procedures based on expectations as summary statistics have overall satisfying performance with globally less sharp posterior distributions. The addition of the posterior log-variances does not seem to significantly change the selected samples.

S4 .5.2 Real observation inversion

We focus on one observation coming from an experiment involving a mineral called Nontronite. The inversion described in Section 6.4 of the manuscript shows the existence of multiple solutions. A complementary test was made to check the relevance of a potential



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

Figure S9: Inversion of a synthetic observation from the Hapke model. The selected samples using four rejection ABC methods are shown, GLLiM-E-ABC expectations in red, GLLiM-EV-ABC in dotted red, semi-automatic ABC in green, GLLiM-L2-ABC in blue and GLLiM-MW2-ABC in black. The margins for $\omega, \bar{\theta}, b$ and c are shown from top to bottom respectively. Columns correspond to different ϵ values, in column from left to right, set to the 1%, 0.1% and 0.05% quantile respectively. The vertical lines indicate the parameter values used for the simulation.

second mode observed for the b parameter in Figure 2 of the manuscript. Figure S10 below, obtained by decreasing the threshold ϵ to the 0.05% quantile, shows that this mode around 0.5 tends to disappear. As an additional check, the reconstructed signal obtained with this value of b was observed to be quite different from the inverted signal (not shown).

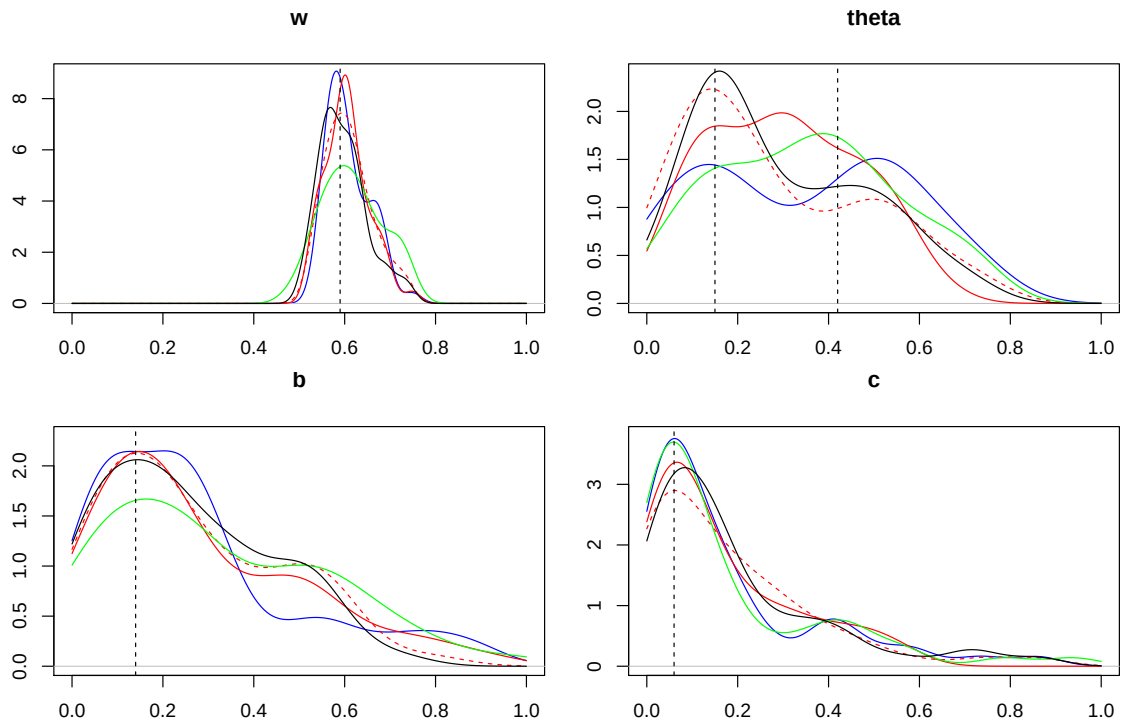


Figure S10: Real observation inversion using the Hapke model. Selected samples using four ABC methods, GLLiM-E-ABC in red, GLLiM-EV-ABC in dotted red, semi-automatic ABC in green, GLLiM-L2-ABC in blue and GLLiM-MW2-ABC in black. The posterior margins for $\omega, \bar{\theta}, b$ and c are shown respectively. The threshold ϵ is set to the 0.05% quantile (50 selected values). The vertical lines indicate the parameters values $(\omega, \bar{\theta}, b, c) = (0.59, 0.15, 0.14, 0.06)$ and $(0.59, 0.42, 0.14, 0.06)$ (identical except for $\bar{\theta}$).

References

Aliprantis, C. D. and K. C. Border (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media.

Bernton, E., P. E. Jacob, M. Gerber, and C. P. Robert (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 81*, 235–269.

Chen, Y., T. T. Georgiou, and A. Tannenbaum (2019). Optimal Transport for Gaussian Mixture Models. *IEEE Access 7*, 6269–6278.

Crackel, R. and J. Flegal (2017). Bayesian inference for a flexible class of bivariate beta distributions. *Journal of Statistical Computation and Simulation 87*, 295–312.

Deleforge, A., F. Forbes, and R. Horaud (2015, September). High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables. *Statistics and Computing 25*(5), 893–911.

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

Geyer, C. J. and L. T. Jonhson (2020). mcmc: Markov chain Monte Carlo. https://cran.r-project.org/web/packages/mcmc/.

Jiang, B., T.-Y. Wu, Z. C., and W. Wong (2017). Learning summary statistics for Approximate Bayesian Computation via Deep Neural Network. *Statistica Sinica*, 1595–1618.

Makarov, B. and A. Podkorytov (2013). *Real analysis: measures, integrals and applications*. Springer Science & Business Media.

Marin, J.-M., P. Pudlo, C. P. Robert, and R. J. Ryder (2012). Approximate Bayesian computation methods. *Statistics and Computing 22*, 1167–1180.

Rakhlin, A., D. Panchenko, and S. Mukherjee (2005). Risk bounds for mixture density estimation. *ESAIM: Probability and Statistics 9*, 220–229.

Schuhmacher, D., B. Bahre, C. Gottschlich, V. Hartmann, F. Heinemann, and B. Schmitzer (2020). transport: Computation of Optimal Transport Plans and Wasserstein Distances. R package version 0.12-2.

Tao, T. (2011). *An introduction to measure theory*. American Mathematical Society Providence, RI.

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Zeevi, A. J. and R. Meir (1997). Density estimation through convex combinations of densities: approximation and estimation bounds. *Neural Networks 10*(1), 99–109.