Approximate Bayesian computation with surrogate posteriors

Florence Forbes, Hien Duy Nguyen, Trung Tin Nguyen, Julyan Arbel

Approximate Bayesian computation with surrogate posteriors

Florence Forbes[1] Hien Duy Nguyen[2] Trung Tin Nguyen[3] Julyan Arbel[1]

[1] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Inria Grenoble Rhone-Alpes, 655 av. de l'Europe, 38335 Montbonnot, France
[2]School of Engineering and Mathematical Sciences, La Trobe University, Bundoora, Australia
[3]Normandie Univ, UNICAEN, CNRS, LMNO, 14000 Caen, France

February 12, 2021

Abstract A key ingredient in approximate Bayesian computation (ABC) procedures is the choice of a discrepancy that describes how different the simulated and observed data are, often based on a set of summary statistics when the data cannot be compared directly. Unless discrepancies and summaries are available from experts or prior knowledge, which seldom occurs, they have to be chosen and this can affect the approximations. Their choice is an active research topic, which has mainly considered data discrepancies requiring samples of observations or distances between summary statistics, to date. In this work, we introduce a preliminary learning step in which surrogate posteriors are built from finite Gaussian mixtures using an inverse regression approach. These surrogate posteriors are then used in place of summary statistics and compared using metrics between distributions in place of data discrepancies. Two such metrics are investigated, a standard L_2 distance and an optimal transport-based distance. The whole procedure can be seen as an extension of the semi-automatic ABC framework to functional summary statistics. The resulting ABC quasi-posterior distribution is shown to converge to the true one, under standard conditions. Performance is illustrated on both synthetic and real data sets, where it is shown that our approach is particularly useful when the posterior is multimodal.

Keywords Approximate Bayesian computation, summary statistics, surrogate models, Gaussian mixtures, discrepancy measures, divergence measures, L_2 distance, Wasserstein distance, multimodal posterior distributions.

Contents

1 Introduction

Approximate Bayesian computation (ABC) (see, *e.g.*, Sisson et al. 2019) appears as a natural candidate for addressing problems where there is a lack of availability of the likelihood. Such cases occur when the direct model or data generating process is not available, analytically, but is available as a simulation procedure; *e.g.*, when the data generating process is characterized as a series of ordinary differential equations, as in Mesejo et al. (2016); Hovorka et al. (2004). In addition, typical features or constraints that can occur in practice are that 1) the observations $\mathbf{y}$ are high-dimensional, because they represent signals in time or spectra, as in Schmidt and Fernando (2015); Bernard-Michel et al. (2009); Ma et al. (2013)

and 2) the parameter $\boldsymbol{\theta}$ to be predicted is itself multi-dimensional with correlated dimensions so that independently predicting its components is sub-optimal; *e.g.*, when there are known constraints such as when the parameter elements are concentrations or probabilities that sum to one (Deleforge et al., 2015a; Lemasson et al., 2016; Bernard-Michel et al., 2009).

ABC enables posterior inference in situations where the likelihood is not available by measuring the similarity between simulated and observed data. The fundamental idea of ABC is to generate parameter proposals $\boldsymbol{\theta}$ in a parameter space Θ using a prior distribution $\pi(\boldsymbol{\theta})$ and accept a proposal if the simulated data $\mathbf{z}$ for that proposal is similar to the observed data $\mathbf{y}$ both in an observation space $\mathcal{Y}$. This similarity is usually measured using a distance or discriminative measure D and a simulated sample $\mathbf{z}$ is retained if $D(\mathbf{z}, \mathbf{y})$ is smaller than a given threshold ϵ. The simulated $\mathbf{z}$ is the result of applying the likelihood black-box to $\boldsymbol{\theta}$. In this simple form, the procedure is generally referred to as rejection ABC. Other variants are possible and often recommended, for instance using MCMC or sequential procedures (*e.g.* Del Moral et al., 2012; Buchholz and Chopin, 2019), but we will focus on the rejection version for the purpose of this paper.

In the case of a rejection algorithm, selected samples are drawn from the so-called ABC quasi-posterior, which is an approximation to the true posterior $\pi(\boldsymbol{\theta} \mid \mathbf{y})$. Under conditions similar to that of Bernton et al. (2019), regarding the existence of a probability density function (pdf) $f_{\boldsymbol{\theta}}(\mathbf{z})$ for the likelihood, the ABC quasi-posterior depends on D and on a threshold ϵ, and can be written as

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, f_{\boldsymbol{\theta}}(\mathbf{z}) \, d\mathbf{z} \, . \tag{1}$$

More specifically, the similarity between $\mathbf{z}$ and $\mathbf{y}$ is generally evaluated based on two components: the choice of summary statistics $s(\cdot)$ to account for the data in a more robust manner, and the choice of a distance to compare the summary statistics. That is, $D(\mathbf{y}, \mathbf{z})$ in (1) should then be replaced by $D(s(\mathbf{y}), s(\mathbf{z}))$, whereupon we abuse the notation and also use D to denote the distance between summary statistics $s(\cdot)$.

However, there is no general rule for constructing good summary statistics for complex models and if a summary statistic does not capture important characteristics of the data, the ABC algorithm is likely to yield samples from an incorrect posterior (Blum et al., 2013; Fearnhead and Prangle, 2012; Gutmann et al., 2018). Great insight has been gained through the work of Fearnhead and Prangle (2012) who introduced the *semi-automatic* ABC framework and showed that under a quadratic loss, the optimal choice for the summary statistic of $\mathbf{y}$ was the true posterior mean of the parameter: $s(\mathbf{y}) = \mathbb{E}[\boldsymbol{\theta} \mid \mathbf{y}]$. This conditional expectation cannot be calculated analytically but can be estimated by regression using a learning data set prior to the ABC procedure itself.

In Fearnhead and Prangle (2012), it is suggested that a simple regression model may be enough to approximate $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{y}]$, but this has since then been contradicted, for instance by Jiang et al. (2017); Wiqvist et al. (2019), who show that the quality of the approximation can matter in practice. Still focusing on posterior means as summary statistics, they use

deep neural networks that capture complex non-linear relationships and exhibit much better results than standard regression approaches. However, deep neural networks remain very computationally costly tools, both in terms of the required size of training data and number of parameters and hyperparameters to be estimated and tuned.

Our first contribution is to investigate an alternative efficient way to construct summary statistics, in the same vein as semi-automatic ABC, but based on posterior moments, not restricted to the posterior means. Although this natural extension was already proposed in Jiang et al. (2017), it requires the availability of a flexible and tractable regression model, able to capture complex non-linear relationships and to provide posterior moments straightforwardly. As such, Jiang et al. (2017) did not consider an implementation of the procedure. For this purpose, the Gaussian Locally Linear Mapping (GLLiM) method (Deleforge et al., 2015b), that we recall in Section 3, appears as a good candidate, with properties in between that of computationally expensive neural networks and standard regression techniques.

In contrast to most regression methods that provide only pointwise predictions, GLLiM provides, at low cost, a parametric estimation of the full true posterior distributions. Using a learning set of parameters and observations couples, GLLiM learns a family of finite Gaussian mixtures whose parameters depend analytically on the observation to be inverted. For any observed data, the true posterior can be approximated as a Gaussian mixture, whose moments are easily computed in closed form and turned into summary statistics for subsequent ABC sample selection.

Moreover, beyond semi-automatic ABC, the obtained parametric approximations of the posterior distributions can be used without reducing them to moments. The idea is to compare directly these surrogate posterior distributions rather than comparing their moments. However, when replacing summary statistics by full surrogate distributions, even parametric ones, the usual distances or discrepancy measures used to compare them must also be changed. Recent developments in optimal transport-based distances designed for Gaussian mixtures (Delon and Desolneux, 2020; Chen et al., 2019) match perfectly this need with a so-called Mixture-Wasserstein distance referred to in Delon and Desolneux (2020) and below as MW_2. Other distances between mixtures are tractable and among them the L_2 distance is also considered in this work, as it is straightforward to compute.

The novelty of our approach and its comparison with existing work is emphasized in Section 2. The GLLiM output is briefly described in Section 3. A first exploitation of GLLiM combined with the semi-automatic ABC principle of Fearnhead and Prangle (2012) is presented in Section 4.1. Our extension using functional summary statistics is then described in Section 4.2. The approach's theoretical properties are investigated in Section 5 and the practical performance is illustrated in Section 6, both on synthetic and real data.

2 Related work

In the works of Nguyen et al. (2020); Jiang et al. (2018); Bernton et al. (2019); Park et al. (2016); Gutmann et al. (2018), the difficulties associated with finding efficient summary statistics was bypassed by adopting, respectively, the Energy Distance, a Kullback–Leibler divergence estimator, the Wasserstein distance, the Maximum Mean Discrepancy (MMD), and classification accuracy to provide a data discrepancy measure. Such approaches compare simulated data and observed data by looking at them as *i.i.d.* samples from distributions, respectively linked to the simulated and true parameter, except for Bernton et al. (2019); Gutmann et al. (2018) that propose solutions to also handle time series. We suspect that to be effective these methods require that the observed and simulated data contain each a moderately large number of samples. Typically, they cannot be applied if we observe only one limited sample related to the parameter to be recovered. This is a major difference with the approach that we propose.

We propose not to compare samples from distributions, but to directly compare the distributions by comparing their surrogates. So doing, we are not concerned with data or sample discrepancies, but with distances between distributions. We can still use the same Wasserstein, Kullback–Leibler divergence, *etc.*, but in their *population* versions rather than in their empirical or estimator versions.

The Wasserstein distance can be computed between Mixtures of Gaussians, thanks to the recent work of Delon and Desolneux (2020); Chen et al. (2019). Note that it is not strictly speaking the Wasserstein distance, but a Wasserstein based distance. Other distances are even simpler to compute. Closed form expressions also exist for the L_2 distance, for the MMD with a Gaussian RBF kernel, or a polynomial kernel (see Sriperumbudur et al., 2010; Muandet et al., 2012) and for the Jensen–Rényi divergence of degree two (see Wang et al., 2009). Kristan et al. (2011) also proposed an algorithm based on the so-called inscented transform in order to compute the Hellinger distance between two Gaussian mixtures, although it is unclear what the complexity of this algorithm is.

In addition, it is always possible to use the previous data discrepancies or their estimators by simulating first samples from the distributions to be compared but this is likely to be computationally sub-optimal. In this work, for illustration, we investigate the use of the L_2 and MW_2 distances. Applications of other distances are left for future research.

3 Parametric posterior approximation with Gaussian mixtures

A learning set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n = [N]\}$ is built from the joint distribution that results from the prior $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ and the likelihood $f_{\boldsymbol{\theta}}$, where $[N] = \{1, \ldots, N\}$. The idea is to capture the relationship between $\boldsymbol{\theta}$ and $\mathbf{y}$ with a joint probabilistic model for which computing conditional distributions and moments is straightforward. For the choice of

the model to fit to $\mathcal{D}_N$, we propose to use the so-called Gaussian Locally Linear Mapping (GLLiM) model (Deleforge et al., 2015b) for its ability to capture non-linear relationships in a tractable manner, based on flexible mixtures of Gaussian distributions. GLLiM can be included in the class of inverse regression approaches, such as sliced inverse regression (Li, 1991), partial least squares (Cook and Forzani, 2019), mixtures of regressions approaches of different variants, *e.g.* mixtures of experts (Nguyen et al., 2019), cluster weighted models (Ingrassia et al., 2012), and kernel methods (Nataraj et al., 2018). In contrast to deep learning approaches (see Arridge et al. 2019 for a survey), GLLiM provides for each observed $\mathbf{y}$, a full posterior probability distribution within a family of parametric models $\{p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}), \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$. To model non-linear relationships, it uses a mixture of K linear models. More specifically, the expression of $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi})$ is analytical and available for all $\mathbf{y}$ with $\boldsymbol{\phi}$ being independent of $\mathbf{y}$:

$$p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}) = \sum_{k=1}^{K} \eta_k(\mathbf{y})\, \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{A}_k\mathbf{y} + \boldsymbol{b}_k, \boldsymbol{\Sigma}_k), \tag{2}$$

where $\mathcal{N}(\,.\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian pdf with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $\eta_k(\mathbf{y}) = \dfrac{\pi_k \mathcal{N}(\mathbf{y}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{y}; \boldsymbol{c}_j, \boldsymbol{\Gamma}_j)}$. This distribution involves a number of parameters $\boldsymbol{\phi} = \{\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k, \boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$. One interesting property of such a parametric model is that the mixture setting provides some guaranties that when choosing K large enough it is possible to approximate any reasonable relationship (Nguyen et al., 2019). The parameter $\boldsymbol{\phi}$ can be estimated by fitting a GLLiM model to the learning set $\mathcal{D}_N$ using a standard Expectation-Maximization (EM) algorithm. Details on the model and its estimation are provided in Deleforge et al. (2015b).

Fitting a GLLiM model to $\mathcal{D}_N$ therefore results in a set of parametric distributions $\{p_G(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\phi}^*_{K,N}),\ \mathbf{y} \in \mathcal{Y}\}$, which are mixtures of Gaussian distributions and can be seen as a parametric mapping from $\mathbf{y}$ values to posterior pdfs on $\boldsymbol{\theta}$. The parameter $\boldsymbol{\phi}^*_{K,N}$ is the same for all conditional distributions and does not need to be re-estimated for each new instance of $\mathbf{y}$. In terms of usage in ABC procedures, the setting is very similar to the ones using standard summary statistics.

Note that when required, a response $\boldsymbol{\theta}$ corresponding to an observed input $\mathbf{y}$ can be proposed using the expectation of $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N})$ in (2) given by:

$$\mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}] = \sum_{k=1}^{K} \eta_k(\mathbf{y})\ (\boldsymbol{A}^*_k\mathbf{y} + \boldsymbol{b}^*_k). \tag{3}$$

It is also straightforward to compute the covariance matrix of $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N})$, which is

given by

$$\mathrm{Var}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}] = \sum_{k=1}^{K} \eta_k(\mathbf{y}) \left[\boldsymbol{\Sigma}^*_k + (\boldsymbol{A}^*_k \mathbf{y} + \boldsymbol{b}^*_k)(\boldsymbol{A}^*_k \mathbf{y} + \boldsymbol{b}^*_k)^\top \right] - \left(\sum_{k=1}^{K} \eta_k(\mathbf{y})(\boldsymbol{A}^*_k \mathbf{y} + \boldsymbol{b}^*_k) \right) \left(\sum_{k=1}^{K} \eta_k(\mathbf{y})(\boldsymbol{A}^*_k \mathbf{y} + \boldsymbol{b}^*_k) \right)^\top \quad (4)$$

where $(\cdot)^\top$ is the matrix transposition operator.

Expression (3) providing approximate posterior means can be directly used in a semi-automatic procedure but, in addition, summary statistics extracted from the covariance expression (4) can also be included and is likely to improve the ABC selection as illustrated in Section 6.

4 Extended semi-automatic ABC

Semi-automatic ABC refers to an approach introduced in Fearnhead and Prangle (2012), which has since then led to various attempts and improvements, see *e.g.* Jiang et al. (2017); Wiqvist et al. (2019) without dramatic deviation from the original ideas.

4.1 Extension to extra summary vectors

In the same vein, a natural idea is to use the approximate posterior expectation provided by GLLiM in (3) as the summary statistic s of data $\mathbf{y}$:

$$s(\mathbf{y}) = \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}],$$

and then to apply standard ABC algorithms, *e.g.* a rejection ABC. This is strictly following the idea of Fearnhead and Prangle (2012) but using a non-standard regression method (GLLiM). It provides a first attempt to combine GLLiM and ABC procedures and has the advantage over neural networks of being easier to estimate without the need of huge learning data sets and obscure hyperparameter tuning.

However, one advantage of GLLiM over most regression methods is not to reduce to pointwise predictions and to provide full posteriors as output. The posteriors can then be used to provide other posterior moments as summary statistics. The same standard ABC procedure as before can be applied but now with $s_1(\mathbf{y}) = \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]$ and $s_2(\mathbf{y}) = \mathrm{Var}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}]$, as given by (4). In Section 6, we show examples where s_2 restricts to the posterior log-variances, *i.e.* the logarithms of the diagonal elements of the posterior covariance matrix.

To summarize, the discussion above leads to the procedure detailed in Algorithm 1. It requires two simulated data sets, one for training GLLiM and constructing the surrogate

posteriors, and one for the ABC selection procedure itself. For rejection ABC, the selection also requires the user to fix a threshold ϵ. It is common practice to set ϵ to a quantile of the computed distances. The use of GLLiM also requires the choice of K, the number of Gaussian components in the mixtures. K can be chosen using model selection criteria (see Deleforge et al., 2015b), but its precise value is not critical all the more so if GLLiM is not used for prediction, directly. See details in our experiments (Section 6).

As illustrated in Section 6, it is easy to construct examples where the posterior expectations even when well-approximated are not performing well as summary statistics. Providing a straightforward and tractable way to add other posterior moments is then already an interesting contribution. However, to really make the most of the GLLiM framework, we propose to further exploit the fact that GLLiM provides more than the mean or variances or other moments. We elaborate further in the next section.

4.2 Extension to functional summary statistics

Instead of comparing simulated $\mathbf{z}$'s to the observed $\mathbf{y}$, or equivalently their summary statistics, we propose to compare the $p_G(\boldsymbol{\theta} \mid \mathbf{z}, \boldsymbol{\phi}^*_{K,N})$'s to $p_G(\boldsymbol{\theta} \mid \mathbf{y}, \boldsymbol{\phi}^*_{K,N})$ as given by (2). As approximation of the true posteriors, these quantities are likely to capture the main characteristics of $\boldsymbol{\theta}$ without committing to the choice of a particular moment. The comparison requires an appropriate distance that needs to be a mathematical distance between distributions. The equivalent functional distance to the L_2 distance can still be used, the Hellinger distance or any other divergence. A natural one is the Kullback–Leibler divergence but computing Kullback–Leibler divergences between mixtures is not straightforward. Computing the Energy statistic (*e.g.* Nguyen et al., 2020) appears at first to be easier but in the end that would still resort to Monte Carlo sums. Since the model (2) is parametric, we could also compute distances between the parameters of the mixtures that depend on $\mathbf{y}$. That is for $k = [K]$, between the $\eta^*_k(\mathbf{y}) = \frac{\pi^*_k \mathcal{N}(\mathbf{y}; \boldsymbol{c}^*_k, \boldsymbol{\Gamma}^*_k)}{\sum_{j=1}^K \pi^*_j \mathcal{N}(\mathbf{y}; \boldsymbol{c}^*_j, \boldsymbol{\Gamma}^*_j)}$ and the $\boldsymbol{A}^*_k\mathbf{y} + \boldsymbol{b}^*_k$'s. But this may lead us back to the usual issue with distances between summary statistics and also we may have to face the label switching issue, not easy to handle within the ABC procedures.

Recently, interesting developments regarding the Wasserstein distance and Gaussian mixtures have emerged (Delon and Desolneux, 2020; Chen et al., 2019), introducing an optimal transport-based distance between Gaussian mixtures. The good properties of this distance make it an interesting candidate for our purpose. We first recall the definition of this distance, denoted by MW_2 and describe our next ABC procedure referred to as GLLiM-MW2-ABC. The L_2 distance between mixtures is also very straightforward to compute and recalled in Section 4.2.2, leading then to another procedure, which we call GLLiM-L2-ABC.

4.2.1 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,

$$\mathrm{MW}_2^2(f_1, f_2) = \min_{\mathbf{w} \in \Pi(\pi_1, \pi_2)} \sum_{k,l} w_{kl}\, \mathrm{W}_2^2(g_{1k}, g_{2l})\,, \tag{5}$$

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 (5) 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 what follows, the MW_2 distances are computed using the **transport** R package.

4.2.2 L_2 distance between Gaussian mixtures

The L_2 distance between two Gaussian distributions g_1 and g_2 is closed form and given by,

$$\mathrm{L}_2(g_1, g_2) = \mathcal{N}(\boldsymbol{\mu}_1; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)\,.$$

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

$$\mathrm{L}_2^2(f_1, f_2) = \sum_{k,l} \pi_{1k}\pi_{2l}\, \mathrm{L}_2^2(g_{1k}, g_{2l})\,, \tag{6}$$

which can be evaluated in $\mathcal{O}(K_1 K_2)$ time. We do not discuss further the different properties of the various possible distances but the distance choice has a potential impact on the associated ABC procedure described in the next section. This impact is illustrated in the experimental Section 6.

4.2.3 Functional GLLiM-ABC

The resulting procedure is very similar to that of Section 4.1. It differs in that no summary statistics are computed per se but distances between surrogate posteriors are compared to a given threshold and used for sample selection. In the next sections, we will often write GLLiM-D-ABC in place of the functional versions GLLiM-MW2-ABC and GLLiM-L2-ABC, to include both cases and possibly other distances D. The semi-automatic ABC extensions that we therefore propose to investigate are all summarized in Algorithm 1.

Algorithm 1 GLLiM-ABC algorithms – Vector and functional variants

1: **Inverse operator learning.** Apply GLLiM on a training set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ to estimate, for any $\mathbf{z} \in \mathcal{Y}$, the K-Gaussian mixture $p_G(\boldsymbol{\theta} \mid \mathbf{z}, \boldsymbol{\phi}^*_{K,N})$ in (2) as a first approximation of the true posterior $\pi(\boldsymbol{\theta} \mid \mathbf{z})$, where $\boldsymbol{\phi}^*_{K,N}$ does not depend on $\mathbf{z}$.

2: **Distances computation.** Consider another simulated set $\mathcal{E}_M = \{(\boldsymbol{\theta}_m, \mathbf{z}_m), m \in [M]\}$. For a given observed $\mathbf{y}$, do one of the following for $m \in [M]$:

   Vector summary statistics. (Section 4.1)

   GLLiM-E-ABC: Compute summary statistics $s_1(\mathbf{z}_m) = \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}]$ (3).

   GLLiM-EV-ABC: Compute both $s_1(\mathbf{z}_m)$ and $s_2(\mathbf{z}_m)$ by considering also posterior log-variances derived from (4).

   In either of these cases, compute standard distances between summary statistics.

   Functional summary statistics. (Section 4.2)

   GLLiM-MW2-ABC: Compute $\mathrm{MW}_2(p_G(\cdot \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}), p_G(\cdot \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}))$.

   GLLiM-L2-ABC: Compute $\mathrm{L}_2(p_G(\cdot \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}), p_G(\cdot \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}))$.

3: **Sample selection.** Select the $\boldsymbol{\theta}_m$ values that correspond to distances under an ϵ threshold, typically the 0.1% distance quantile (rejection ABC) or apply some standard ABC procedure that can handle distances, directly.

4: **Sample use.** For a given observed $\mathbf{y}$, use the produced sample of $\boldsymbol{\theta}$ values to compute a closer approximation of $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ or to obtain a better prediction using the empirical mean of the retained sample.

5 Theoretical properties

Before illustrating the proposed GLLiM-D-ABC procedures performance, we investigate the theoretical properties of our ABC quasi-posterior defined via surrogate posteriors.

Let $\mathcal{X} = \Theta \times \mathcal{Y}$ and $(\mathcal{X}, \mathcal{F})$ be a measurable space. Let λ be a σ-finite measure on $\mathcal{F}$. Whenever we mention below that a probability measure Pr on $\mathcal{F}$ has a density, we will understand that it has a Radon–Nikodym derivative with respect to λ (λ can typically be chosen as the Lebesgue measure on the Euclidean space). For all $p \in [1, \infty)$ and f, g in appropriate spaces, let $D_p(f,g) = \left(\int |f(\mathbf{x}) - g(\mathbf{x})|^p \, d\lambda(\mathbf{x})\right)^{1/p}$ denote the L_p distance and $D_H^2(f,g) = \int(\sqrt{f(\mathbf{x})} - \sqrt{g(\mathbf{x})})^2 d\lambda(\mathbf{x})$ be the squared Hellinger distance. When not specified otherwise, let D be an arbitrary distance on $\mathcal{Y}$ or on densities, depending on the context. We further denote the L_p norm for vectors by $\|\cdot\|_p$.

In a GLLiM-D-ABC procedure, the ABC quasi-posterior is constructed as follows. Let $p_G^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) = p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N})$ be the surrogate conditional distribution of form (2), learned from a preliminary GLLiM model with K components and using a learning set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n = [N]\}$. This conditional distribution is a K-component mixture, which depends on a set of learned parameters $\boldsymbol{\phi}^*_{K,N}$, independent of $\mathbf{y}$. The GLLIM-D-ABC quasi-posterior resulting from the GLLiM-D-ABC procedure then depends both on K, N and the tolerance level ϵ and can be written as

$$q_{G,\epsilon}^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(p_G^{K,N}(\,.\,|\mathbf{y}),\, p_G^{K,N}(\,.\,|\mathbf{z})) \leq \epsilon\}} \, f_{\boldsymbol{\theta}}(\mathbf{z}) \, d\mathbf{z} \,, \tag{7}$$

where D is a distance on densities such as the MW_2 and L_2 metrics which are both proper distances as recalled, previously.

We provide two types of results, below. In the first result (Theorem 1), the true posterior is used to compare samples $\mathbf{y}$ and $\mathbf{z}$. This result aims at providing insights on the proposed quasi-posterior formulation and at illustrating its potential advantages. In the second result (Theorem 2), a surrogate posterior is learned and used to compare samples. Conditions are specified under which the resulting ABC quasi-posterior converges to the true posterior.

5.1 Convergence of the ABC quasi-posterior

In this section, we assume a fixed given observed $\mathbf{y}$ and the dependence on $\mathbf{y}$ is omitted from the notation when there is no confusion.

Let us first recall the standard form of the ABC quasi-posterior, omitting summary statistics from the notation:

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, f_{\boldsymbol{\theta}}(\mathbf{z}) \, d\mathbf{z} \,. \tag{8}$$

If D is a distance and $D(\mathbf{y}, \mathbf{z})$ is continuous in $\mathbf{z}$, the ABC posterior in (8) can be shown to have the desirable property of converging to the true posterior when ϵ tends to 0 (see Prangle et al., 2018).

The proof is based on the fact that when ϵ tends to 0, due to the property of the distance D, the set $\{\mathbf{z} \in \mathcal{Y} : D(\mathbf{y}, \mathbf{z}) \leq \epsilon\}$, defining the indicator function in (8), tends to the singleton $\mathbf{y}$ so that consequently $\mathbf{z}$ in the likelihood can be replaced by the observed $\mathbf{y}$, which then leads to an ABC quasi-posterior proportional to $\pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\mathbf{y})$ and therefore to the true posterior as desired (see also Rubio and Johansen, 2013; Bernton et al., 2019). It is interesting to note that this proof is based on working on the term under the integral only and is using the equality, at convergence, of $\mathbf{z}$ to $\mathbf{y}$, which is actually a stronger than necessary assumption for the result to hold. Alternatively, if we first rewrite (8) using Bayes' theorem, it follows that

$$\begin{aligned} \pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) &\propto \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}}\ \pi(\boldsymbol{\theta})\ f_{\boldsymbol{\theta}}(\mathbf{z})\ d\mathbf{z} \\ &\propto \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}}\ \pi(\boldsymbol{\theta} \mid \mathbf{z})\ \pi(\mathbf{z})\ d\mathbf{z}\ . \end{aligned} \tag{9}$$

That is, when accounting for the normalizing constant:

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}}\ \pi(\boldsymbol{\theta} \mid \mathbf{z})\ \pi(\mathbf{z})\ d\mathbf{z}}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}}\ \pi(\mathbf{z})\ d\mathbf{z}}\ . \tag{10}$$

Using this equivalent formulation, we can then replace $D(\mathbf{y}, \mathbf{z})$ by $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$, with D now denoting a distance on densities, and obtain the same convergence result when ϵ tends to 0. More specifically, we can show the following general result. Let us define our ABC quasi-posterior as,

$$q_\epsilon\left(\boldsymbol{\theta} \mid \mathbf{y}\right) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}}\ f_{\boldsymbol{\theta}}(\mathbf{z})\ d\mathbf{z}, \tag{11}$$

which can be written as

$$q_\epsilon\left(\boldsymbol{\theta} \mid \mathbf{y}\right) = \frac{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) \pi\left(\mathbf{z}\right) d\mathbf{z}}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}),\pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi\left(\mathbf{z}\right) d\mathbf{z}}\ . \tag{12}$$

The following theorem shows that $q_\epsilon\left(\cdot \mid \mathbf{y}\right)$ converges to $\pi\left(\cdot \mid \mathbf{y}\right)$ in total variation, for fixed $\mathbf{y}$. The proof is detailed in Subsection 8.1.

Theorem 1. *For every $\epsilon > 0$, let $A_\epsilon = \{\mathbf{z} \in \mathcal{Y} : D\left(\pi\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{z}\right)\right) \leq \epsilon\}$. Assume the following:*

(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), $q_\epsilon\left(\cdot \mid \mathbf{y}\right)$ *in (12) converges in total variation to* $\pi\left(\cdot \mid \mathbf{y}\right)$*, for fixed* $\mathbf{y}$*, as* $\epsilon \to 0$.

It appears that what is important is not to select $\mathbf{z}$'s that are close (and at the limit equal) to the observed $\mathbf{y}$ but to choose $\mathbf{z}$'s so that the posterior $\pi(\,\cdot \mid \mathbf{z})$ (the term appearing in the integral in (9)) is close (and at the limit equal) to $\pi(\,\cdot \mid \mathbf{y})$. And this last property is less demanding than $\mathbf{z} = \mathbf{y}$. Potentially, there may be several $\mathbf{z}$'s satisfying $\pi(\,\cdot \mid \mathbf{z}) = \pi(\,\cdot \mid \mathbf{y})$, but this is not problematic when using (9), while it is problematic when following the standard proof as in Bernton et al. (2019).

5.2 Convergence of the ABC quasi-posterior with surrogate posteriors

In most ABC settings based on data discrepancy or summary statistics, the above consideration and result are not useful because the true posterior is unknown by construction and cannot be used to compare samples. However this principle becomes useful in our setting, which is based on surrogate posteriors. While the previous result can be seen as an oracle of sort, it is more interesting in practice to investigate whether a similar result holds when using surrogate posteriors in the ABC likelihood. This is the goal of Theorem 2 below, which we prove for a restricted class of target distribution and of surrogate posteriors that are learned as mixtures.

We now assume that $\mathcal{X} = \Theta \times \mathcal{Y}$ is a compact set and consider the following class $\mathcal{H}_\mathcal{X}$ of distributions on $\mathcal{X}$, $\mathcal{H}_\mathcal{X} = \{g_{\boldsymbol{\varphi}} : \boldsymbol{\varphi} \in \Psi\}$, with constraints on the parameters, Ψ being a bounded parameter set. In addition the densities in $\mathcal{H}_\mathcal{X}$ are assumed to satisfy for any $\boldsymbol{\varphi}, \boldsymbol{\varphi}' \in \Psi$,

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

where a, b and B are arbitrary positive scalars.

We denote by p^K a K-component mixture of distributions from $\mathcal{H}_\mathcal{X}$ and defined for all $\mathbf{y} \in \mathcal{Y}$, $p^{K,N}\left(\cdot \mid \mathbf{y}\right)$ as follows:

$$\forall \boldsymbol{\theta} \in \Theta, \quad p^{K,N}\left(\boldsymbol{\theta} \mid \mathbf{y}\right) = p^K\left(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}\right),$$

with $\boldsymbol{\phi}^*_{K,N}$ the maximum likelihood estimate (MLE) for the data set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ generated from the true joint distribution $\pi(\cdot, \cdot)$:

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

In addition, for every $\epsilon > 0$, let $A_{\epsilon,\mathbf{y}}^{K,N} = \left\{\mathbf{z} \in \mathcal{Y} : D\left(p^{K,N}\left(\cdot \mid \mathbf{y}\right), p^{K,N}\left(\cdot \mid \mathbf{z}\right)\right) \leq \epsilon\right\}$ and $q_\epsilon^{K,N}$ denote the ABC quasi-posterior defined with $p^{K,N}$ by

$$q_\epsilon^{K,N}\left(\boldsymbol{\theta} \mid \mathbf{y}\right) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}),p^{K,N}(\cdot|\mathbf{z}))\leq\epsilon\}}\, f_{\boldsymbol{\theta}}(\mathbf{z})\, d\mathbf{z}\,. \tag{13}$$

Theorem 2. *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) $D\left(\cdot,\cdot\right) : \Pi \times \Pi \to \mathbb{R}_+ \cup \{0\}$ *is a metric on a functional class* Π*, which contains the class*

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

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

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

Then, under (B1)–(B4), the Hellinger distance $D_{\mathrm{H}}\left(q_\epsilon^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right)$ *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*

$$\lim_{\epsilon\to 0, K\to\infty, N\to\infty} \Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) \geq \beta\right\}\right) \leq \alpha\right) = 1. \tag{14}$$

Sketch of the proof of Theorem 2. For all $\boldsymbol{\theta} \in \Theta, \mathbf{y} \in \mathcal{Y}$, the quasi-posterior (13) can be written equivalently as

$$q_\epsilon^{K,N}\left(\boldsymbol{\theta} \mid \mathbf{y}\right) = \int_{\mathcal{Y}} K_\epsilon^{K,N}\left(\mathbf{z}; \mathbf{y}\right) \pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) d\mathbf{z},$$
$$\text{with } K_\epsilon^{K,N}\left(\mathbf{z}; \mathbf{y}\right) = \frac{\mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}),p^{K,N}(\cdot|\mathbf{z}))\leq\epsilon\}}\pi\left(\mathbf{z}\right)}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}),p^{K,N}(\cdot|\mathbf{z}))\leq\epsilon\}}\pi\left(\mathbf{z}\right) d\mathbf{z}},$$

where $K_\epsilon^{K,N}\left(\cdot; \mathbf{y}\right)$ is a pdf, with respect to $\mathbf{z} \in \mathcal{Y}$, with compact support $A_{\epsilon,\mathbf{y}}^{K,N} \subset \mathcal{Y}$, by definition of $A_{\epsilon,\mathbf{y}}^{K,N}$ and (B4). Using the relationship between Hellinger and L_1 distances (see details in Section 8.2 relations (31) and (32)), it then holds that

$$D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) \leq 2 D_{\mathrm{H}}\left(\pi(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi\left(\cdot \mid \mathbf{y}\right)\right), \tag{15}$$

where there exists $\mathbf{z}_{\epsilon,\mathbf{y}}^{K,N} \in B_{\epsilon,\mathbf{y}}^{K,N}$ with

$$B_{\epsilon,\mathbf{y}}^{K,N} = \underset{\mathbf{z}\in A_{\epsilon,\mathbf{y}}^{K,N}}{\arg\max}\, D_1\left(\pi\left(\cdot \mid \mathbf{z}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right).$$

The next step is to bound the right-hand side of (15) using the triangle inequality with respect to the Hellinger distance D_{H}. 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 $\epsilon > 0$, $\mathbf{z}_{\epsilon,\mathbf{y}}^{K,N} \in A_{\epsilon,\mathbf{y}}^{K,N}$ it holds that $\mathbf{z}_{0,\mathbf{y}}^{K,N} \in A_{0,\mathbf{y}}^{K,N}$, 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 (15) can then be decomposed in three parts,

$$\begin{aligned} D_{\mathrm{H}}\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) \leq\; & D_{\mathrm{H}}\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) + D_{\mathrm{H}}\left(\pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}\left(\cdot \mid \mathbf{y}\right)\right) \\ & + D_{\mathrm{H}}\left(p^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) . \end{aligned} \tag{16}$$

The first term in 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 $p^{K,N}\left(\cdot \mid \mathbf{y}\right) = p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})$.

Using the fact that $\pi(\cdot \mid \cdot)$ is a uniformly continuous function in $(\boldsymbol{\theta}, \mathbf{y})$ on a compact set $\mathcal{X}$ and taking the limit $\epsilon \to 0$, yields $\lim_{\epsilon\to 0} D_H^2\left(\pi(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) = 0$ in measure λ, with respect to $\mathbf{y} \in \mathcal{Y}$. Since this result is true whatever the data set $\mathcal{D}_N$, it also holds in probability with respect to $\mathcal{D}_N$. That is, given any $\alpha_1 > 0$, $\beta_1 > 0$, there exists $\epsilon\left(\alpha_1, \beta_1\right) > 0$ such that for any $0 < \epsilon < \epsilon\left(\alpha_1, \beta_1\right)$,

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

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 $\mathcal{D}_N$. Such convergences are obtained thanks to Rakhlin et al. (2005, Corollary 2.2), and Lemma 2, which provides the guarantee that we can choose a measurable function $\mathbf{y} \mapsto \mathbf{z}_{0,\mathbf{y}}^{K,N}$. Equation (14) in Theorem 2 follows from the triangular inequality (16). A more technical detailed proof is provided in Subsection 8.2.

Remark. The GLLiM model involving multivariate unconstrained Gaussian distributions does not satisfy the conditions of Theorem 2 so that $p^{K,N}$ cannot be replaced by $p_G^{K,N}$ in the theorem. However as illustrated in Rakhlin et al. (2005), truncated Gaussian distributions with constrained parameters can meet the restrictions imposed in the theorem. We are not aware of any more general result involving the MLE of Gaussian mixtures. The GLLiM model could as well be replaced by another model satisfying the conditions of the theorem but for practical applications, this model would need to have computational properties such as the tractability of the estimation of its parameters and needs to be efficient in multivariate and potentially high-dimensional settings.

6 Numerical experiments

Most benchmark examples in ABC correspond to unimodal and light tailed posterior distributions. Such settings may not be the most appropriate to show differences and discriminate between methods performance. We therefore consider settings that are simple in terms of dimension and complexity but exhibit posterior distributions with characteristics such as bimodality and heavy tails. A first set of three synthetic examples are considered with parameters in dimensions 1 or 2 and bimodal posterior distributions (Section 6.1). A fourth example is derived from a real application in sound source localization where the posterior distribution has mass on two 1D manifolds (Section 6.2). All of these examples are run for a single observation in $d = 10$ dimensions. This choice of dimension is relatively low but corresponds to the dimensions met in practice in some targeted real applications. In particular, we are interested in a real remote sensing inverse problem in planetary science, which is illustrated in Section 6.3.

In this section, the performance of the proposed approaches is assessed and compared. The most sophisticated ABC procedures are not considered as our main focus is on an appropriate choice of summary statistics. All reported results are obtained with a simple rejection scheme as per instances implemented in the **abc** R package (Csillery et al., 2012). The other schemes available in the **abc** package have been tested but no notable performance differences were observed. For comparison we consider the semi-automatic ABC implementation of Fearnhead and Prangle (2012).

To circumvent the choice of an arbitrary summary statistic, Fearnhead and Prangle (2012) showed that the best summary statistic, in terms of the minimal quadratic loss, is the posterior mean. This posterior mean is not known and needs to be approximated. In Fearnhead and Prangle (2012) a regression approach is proposed to provide a way to compute summary statistics prior to the ABC rejection sampling, itself. In this paper, the transformations used for the regression part are $(1, y, y^2, y^3, y^4)$ following the procedure suggested in the **abctools** package (Nunes and Prangle, 2015). We refer to this procedure as semi-automatic ABC. We did not try to optimise the procedure using other transformations but did not notice systematic improvements when increasing the number of polynomial terms, for instance. This approach using the posterior mean is further developed in Jiang et al. (2017), where a MLP deep neural network regression model is employed and replaces the linear regression model of Fearnhead and Prangle (2012). The deep neuronal network with multiple hidden layers considered by Jiang et al. (2017) offers stronger representational power to approximate the posterior mean and hence to learn an informative summary statistic, when compared to linear regression models. Improved results were obtained by Jiang et al. (2017), but we did not compare our approach to their method. As our current examples are of relatively small dimension d, we did not compare either with discrepancy-based ABC techniques such as WABC (Bernton et al., 2019) or classification ABC (Gutmann et al., 2018).

We provide a comparison with a rejection ABC, where we use GLLiM to compute the

posterior expectations that are used as summary statistics. We refer to this procedure as GLLiM-E-ABC. We then augment the summary statistics with approximations of the posterior log-variances, obtained from GLLiM. This method is referred to as GLLiM-EV-ABC. Lastly, the other distance-based procedures are designated as GLLiM-MW2-ABC and GLLiM-L2-ABC.

The procedures differ in the way the distances between each simulation and the observed data $\mathbf{y}$ are defined and computed. In regards to the final sample thresholding (*i.e.*, choice of ϵ), following common practice, all methods retain samples for which the distance to the observation is under a small (*e.g.* 0.1%) quantile of all computed distances.

In all our examples, the dimension of the observed $\mathbf{y}$ is $d = 10$. A GLLiM model with K components and isotropic covariances ($\mathbf{\Sigma}_k$, $k \in [K]$) is learned on a set $\mathcal{D}_N$ of N simulations from the true model. Another set of simulated couples $(\boldsymbol{\theta}, \mathbf{y})$ of size M is used for the ABC rejection scheme. The isotropic GLLiM is simpler than the fully-specified GLLiM and is consistent with the fact that the dimensions of the $\mathbf{y}$'s in our synthetic examples are uncorrelated by design. In more general cases, this simple isotropic GLLiM may also provide surrogate posteriors of sufficient quality for the ABC selection scheme. The GLLiM model is learned using the R package **xLLiM** available on the CRAN (Perthame et al., 2017).

6.1 Non-identifiable models

It is straightforward to construct models that lead to multimodal posteriors by considering likelihoods that are invariant by some transformation.

6.1.1 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

$$\mathbf{y} = F(\boldsymbol{\theta}) + \boldsymbol{\varepsilon} \tag{17}$$

where F is a deterministic theoretical model coming from experts and ε 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 centered Gaussian noise. Non-identifiability may then come when $F(-\boldsymbol{\theta}) = F(\boldsymbol{\theta})$. Following this generative approach, a first simple example is constructed with a Student distributed noise leading to the likelihood:

$$f_{\boldsymbol{\theta}}(\mathbf{y}) = \mathcal{S}_d(\mathbf{y}; \mu^2 \mathbb{I}_d, \sigma^2 \mathrm{I}_d, \nu), \tag{18}$$

where $\mathcal{S}_d(\cdot; \mu^2 \mathbb{1}_d, \sigma^2 \mathrm{I}_d, \nu)$ is the pdf of a d-variate Student 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 distribution, a diagonal scale matrix is not inducing independent dimensions so that $\mathbf{y}$ is not a set of *i.i.d.* univariate Student 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 we set the tolerance level ϵ to the distance 0.1% quantile, so that all selected posterior samples are of size 100. To visualize the densities of posterior samples, we use a density estimation procedure based on the **ggplot2** R package with a Gaussian kernel.

Figure 1 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 in red) and the semi-automatic ABC procedure (in green) both show over dispersed samples with wrongly located modes. The GLLiM-EV-ABC (dotted red line) exhibits two well located modes but does not preserve the symmetry of the true posterior. The distance-based approaches, GLLiM-L2-ABC (blue) and GLLiM-MW2-ABC (black) 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.

In the following subsections we present two further cases that cannot be cast as the above generating process but also exhibit a transformation invariant likelihood.

6.1.2 Sum of moving average models of order 1 (MA(1))

Moving average (MA) models are commonly studied in the ABC literature, see *e.g.* Marin et al. (2012); Jiang et al. (2018); Nguyen et al. (2020). The MA(1) process is a stochastic process $(y'_t)_{t \in \mathbb{N}^*}$ defined by

$$y'_t = z_t + \rho z_{t-1} \, . \tag{19}$$

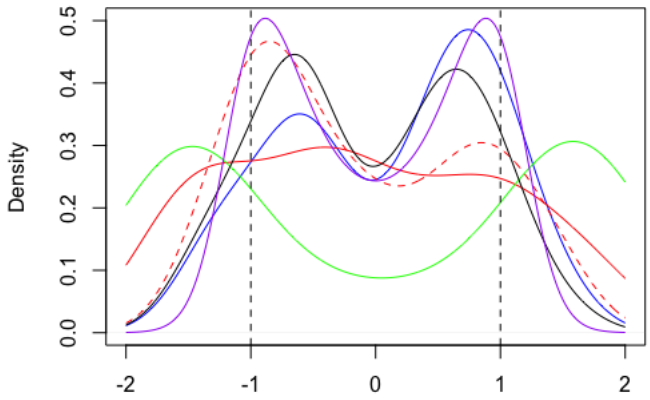


Figure 1: Non identifiable Student distribution. ABC posterior distributions from the selected samples. GLLiM-L2-ABC in blue, GLLiM-MW2-ABC in black, semi-automatic ABC in green, GLLiM-E-ABC (expectations) in red and GLLiM-EV-ABC (expectations and log-variances) in dotted red line. The true posterior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

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

$$y'_t = z_t + \rho z_{t-1} \tag{20}$$
$$y''_t = z'_t - \rho z'_{t-1} \tag{21}$$
$$y_t = y'_t + y''_t \tag{22}$$

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 be then chosen so as to exhibit two symmetric modes. The true posterior looks similar to the previous one but ρ is now a parameter impacting the variance of the likelihood.

For all procedures, we set $N = M = 10^5$, and ϵ to the 0.1% distance quantile so that all selected posterior samples are of size 100. In terms of difficulty, the main difference with the previous example lies in a higher non-linearity of the likelihood and of the model joint distribution. We then report results with a higher $K = 20$. 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 2. 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.

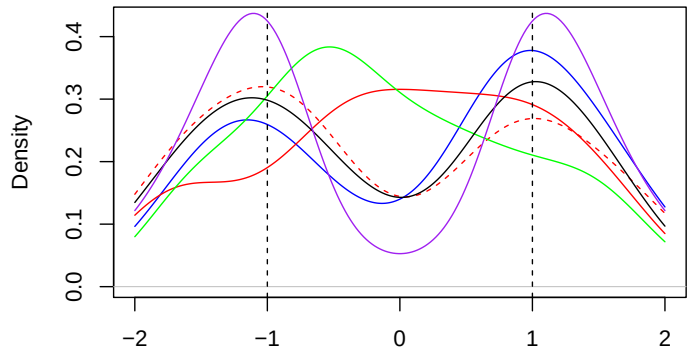


Figure 2: Sum of MA(1) models. ABC posterior distributions from the selected samples. GLLiM-L2-ABC in blue, GLLiM-MW2-ABC in black, semi-automatic ABC in green, GLLiM-E-ABC (expectations) in red and GLLiM-EV-ABC (expectations and log-variances) in dotted red line. The true posterior is shown in purple. The dashed lines indicate the ρ (equivalent) values used to generate the observation.

6.1.3 Sum of moving average models of order 2 (MA(2))

The same principle as in the previous section can be applied to create bimodal posterior distributions from MA(2) processes. The MA(2) process is a stochastic process $(y'_t)_{t\in\mathbb{N}^*}$ defined by

$$y'_t = z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2}, \tag{23}$$

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 uniform on the triangle described by the inequalities

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

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

$$y'_t = z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2} \tag{24}$$
$$y''_t = z'_t - \theta_1 z'_{t-1} + \theta_2 z'_{t-2} \tag{25}$$
$$y_t = y'_t + y''_t, \tag{26}$$

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 we set 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 $\theta_1 = 1$ and $\theta_2 = 0.6$. ABC posterior distribution estimates are shown in Figure 3.

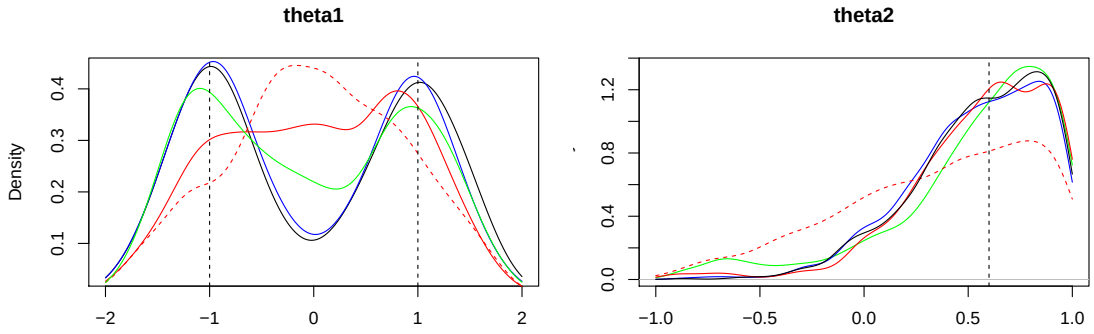


Figure 3: Posterior marginals from the samples selected with a 1% quantile (1000 values): semi-automatic ABC (green), GLLiM-L2-ABC (blue), GLLiM-MW2-ABC (black), GLLiM-E-ABC (red) and GLLiM-EV-ABC (red dot). The dashed lines show the values used to simulate the observation $\theta_1 = 1$ and $\theta_2 = 0.6$.

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 in this paper, none of the considered ABC procedures is fully satisfactory, in that the selected samples are all quite dispersed when compared to the true posterior. This is mainly due to the relatively low size of the observation ($d = 10$). This can be also 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 $d = 50$, and this 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 3, provide an interesting comparison. ABC-D-GLLiM procedures show more 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 does not capture well the bimodality on θ_1 and the addition of the posterior log-variances in GLLiM-EV-ABC does not appear to improve on GLLiM-E-ABC, as observed in the MA(1) case with $K = 10$. In contrast, the semi-automatic ABC procedure shows a bimodal distribution on θ_1 but appears to place too much mass around $\theta_1 = 0$ and $\theta_2 = -1$. These results suggest that although GLLiM in this case may not provide good approximations of the first posterior moments in comparison to semi-automatic ABC, it can still provide good enough approximations of the surrogate posteriors in GLLiM-D-ABC.

6.2 Sound source localization

The next example is constructed from a real sound source localization problem in audio processing. Although microphone arrays provide the most accurate sound source localization, setups limited to two microphones, *e.g.* Beal et al. (2003); Hospedales and Vijayakumar (2008), are often considered to mimic binaural hearing that resembles the real head with applications such as autonomous humanoid robot modelling. Binaural localization cues (Wang and Brown, 2006) include interaural time difference (ITD), interaural level difference (ILD) and interaural phase difference (IPD).

Here we consider an artificial two microphone setup in a 2D scene. The object of interest is a sound source located at an unknown position $\boldsymbol{\theta} = (x, y)$. The two microphones are assumed to be located at known positions, respectively denoted by $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$. A good cue for the sound source localization is the interaural time difference (ITD). The ITD is the difference between two times: the time a sound emitted from the source is acquired by microphone 1 at $\boldsymbol{m}_1$ and the time at microphone 2 at $\boldsymbol{m}_2$. ITD values are widely used by auditory scene analysis methods (Wang and Brown, 2006).

The function F that maps a location $\boldsymbol{\theta}$ onto an ITD observation is

$$F(\boldsymbol{\theta}) = \frac{1}{c}(\|\boldsymbol{\theta} - \boldsymbol{m}_1\|_2 - \|\boldsymbol{\theta} - \boldsymbol{m}_2\|_2), \tag{27}$$

where c is the sound speed in real applications but set to 1 in our example for the purpose of illustration. The important point is that an ITD value does not correspond to a unique point in the scene space, but rather to a whole surface of points. In fact, each isosurface defined by (27) is represented by one sheet of a two-sheet hyperboloid in 2D. Hence, each ITD observation constrains the location of the auditory source to lie on a 1D manifold. The corresponding hyperboloid is determined by the sign of the ITD. In our example, to create a bimodal posterior, we therefore modify the usual setting by taking the absolute value of the ITD so that solutions can now lie on either of the two hyperboloids. In

addition we assume that ITDs are observed with some Student noise that implies heavy tails and possible outliers. Although the ITD is a univariate measure, we consider a more general d dimensional setting by defining the following Student likelihood, with $d = 10$, $\mathbf{y} = (y_1, \ldots, y_d)$ and $\text{ITD}(\boldsymbol{\theta}) = \mid \|\boldsymbol{\theta} - \boldsymbol{m}_1\|_2 - \|\boldsymbol{\theta} - \boldsymbol{m}_2\|_2 \mid$, where

$$f_{\boldsymbol{\theta}}(\mathbf{y}) = \mathcal{S}_d(\mathbf{y}; \text{ITD}(\boldsymbol{\theta})\mathbb{1}_d, \sigma^2 \mathrm{I}_d, \nu) \; . \tag{28}$$

With $d = 10$, the above likelihood corresponds to a 10-variate Student distribution with a 10-dimensional location parameter with all dimensions equal to $\text{ITD}(\boldsymbol{\theta})$, diagonal isotropic scale matrix equal to $\sigma^2 \, \mathrm{I}_d$ and degree-of-freedom (dof) parameter ν.

The parameter space is assumed to be $\Theta = [-2, 2] \times [-2, 2]$ and the prior on $\boldsymbol{\theta}$ is assumed to be uniform on Θ. The microphones positions are $\boldsymbol{m}_1 = (-1, 0)$ and $\boldsymbol{m}_2 = (1, 1)$. We assume $\nu = 1$ and $\sigma^2 = 0.01$. The true $\boldsymbol{\theta}$ is set to $\boldsymbol{\theta} = (0.6, 1)$ and we simulate a 10-dimensional $\mathbf{y}$ following model (28).

We compare the four ABC methods using GLLiM. First, to learn a GLLiM model representation, a training set of $N = 10^5$ pairs $(\boldsymbol{\theta}, \mathbf{y}) \in \Theta \times \mathbb{R}^{10}$ is simulated from a uniform distribution on Θ and by applying model (28). The GLLiM model used consists of $K = 20$ Gaussian components with isotropic covariances. To run the ABC procedures, another training set is simulated with $M = 10^6$. A selected set of 1000 samples is retained by thresholding the distances under the 0.1% quantile.

Figure 4 shows the ABC samples with another sample simulated from the GLLiM posterior distribution, corresponding to the observation $\mathbf{y}$ (Figure 4 (d)). This GLLiM posterior is a 20-component Gaussian mixture of form (2). Another sample obtained using the Metropolis–Hastings algorithm, as implemented in the R package **mcmc** (Geyer and Jonhson, 2020), is shown in Figure 5 (a)). The Metropolis–Hastings sample and the contours of the true posterior (Figure 5 (b)) show that the true posterior concentrates quite sharply around the hyperboloids, which are symmetric with respect to the microphones line and its mediatrice, and contains the true sound source localization as expected.

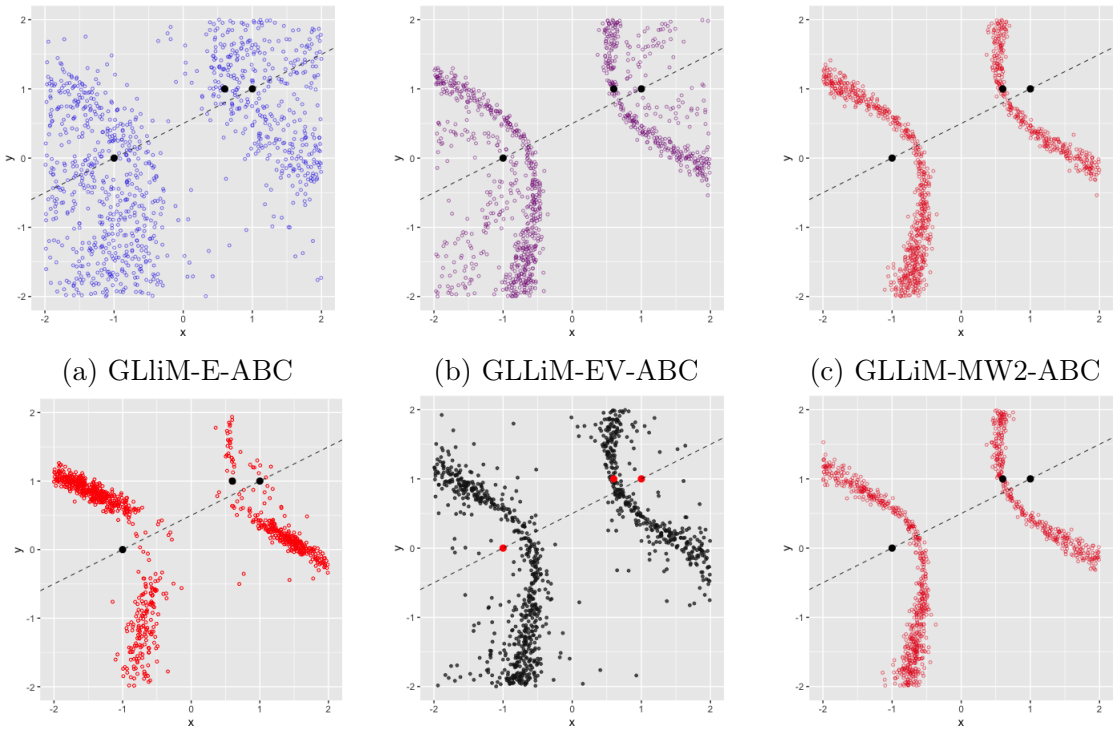
All tested procedures reflect the bimodality of the posterior distribution. The 20-component GLLiM mixture (Figure 4 (d)) reproduces correctly the bimodality of the true posterior. However, the accuracy is clearly improved when using an additional ABC step, with GLLiM-D-ABC. The GLLiM-L2-ABC and GLLiM-MW2-ABC lead to very similar selected samples (Figure 4 (c,f)). The difference between the two distances is better seen in Figure 5 (c,d). These plots show the MW_2 and L_2 distances before selection, *i.e.* for parameters values sampled uniformly in $[-2, 2] \times [-2, 2]$, with large (resp. small) distances colored in red (resp. in blue). The L_2 distance appears to produce sharper differences around the hyperboloids but appear spatially nosier, while MW_2 produces a more homogeneous map, suggesting robustness of this metric to small variations in the posterior distributions. In contrast, using only the GLLiM posterior expectations as summary statistics is not informative enough although the GLLiM mixture itself appears as a reasonable approximation that well captures the main shape of the true posterior. Adding the log-variances to the summary statistics improves the selected sample but it remains too dispersed away from

the modal hyperboloids (Figure 4 (b)). Interestingly, the semi-automatic ABC procedure provides a sample that rather well locates the modal hyperboloids (Figure 4 (e)). As with GLLiM-E-ABC, the procedure is based on a preliminary estimation of the posterior means but using a standard linear regression approach on transformations of the data (here, using monomials up to order 4). In our previous implementations of semi-automatic ABC, the regression part uses the larger data set of size M, while GLLiM uses one of size N. Since, in this example M is 10 times larger than N, to make the comparison with GLLiM more fair, the data set used in the regression part of semi-automatic ABC has been reduced to a size of N, while the set for the ABC step is maintained at size M. Comparisons of the selected samples suggests the superiority of this regression method over GLLiM, with $K = 20$, in capturing non-linearities and in estimating the posterior means. Although GLLiM approximations of the moments may not compare well, the results are greatly improved when using the full GLLiM posteriors as summary statistics.

6.3 A physical model inversion in planetary science

As a real-world example, we consider a remote sensing application coming from the study of planetary environment, in particular the morphological, compositional, photometrical and textural characterization of sites on the surface of a planet. The composition of the surface materials is generally established on the basis of spectral mixing and physical modelling techniques using images produced by hyperspectral cameras, from different angles during a site flyover. An example for the Mars planet is described in (Murchie et al., 2009; Fernando et al., 2016). Such observations can also be measured in the laboratory, on known materials in order to validate a model. In both cases, the interpretation of the surface Bidirectional Reflectance Distribution Factor (BRDF) extracted from these observations is based on the inversion of a physical model of radiative transfer, linking physical and observable parameters in a non-linear way.

The Hapke model is a semi-empirical photometric model that relates physically meaningful parameters to the reflectivity of a granular material for a given geometry of illumination and viewing. Formally, it links a set of parameters $\boldsymbol{\theta} \in \mathbb{R}^4$ to a *theoretical* BRDF denoted by $\mathbf{y} = F_{\text{Hapke}}(\boldsymbol{\theta}) \in \mathbb{R}^d$. A given experiment defines d geometries of measurement, each parameterized by a triplet (θ_0, θ, ϕ) of incidence, emergence and azimuth angles. Moreover, $\boldsymbol{\theta} = (\omega, \overline{\theta}, b, c)$ are the sensitive parameters, respectively single scattering albedo, macroscopic roughness, asymmetry parameter and backscattering fraction. More details on these quantities and their photometric meanings may be found for example in Schmidt and Fernando (2015); Labarre (2017), alongside the explicit expression of F_{Hapke}. Although available, the expression of F_{Hapke} is very complex and tedious to handle, analytically, with a number of approximations required (see for instance the description of the function in more than 15 pages in Labarre 2017). In practice, it is therefore mainly used via a numerical code, allowing simulations from the model. In addition, previous studies, *e.g.* in Kugler et al. (2020); Schmidt and Fernando (2015), have shown evidence for the existence of the



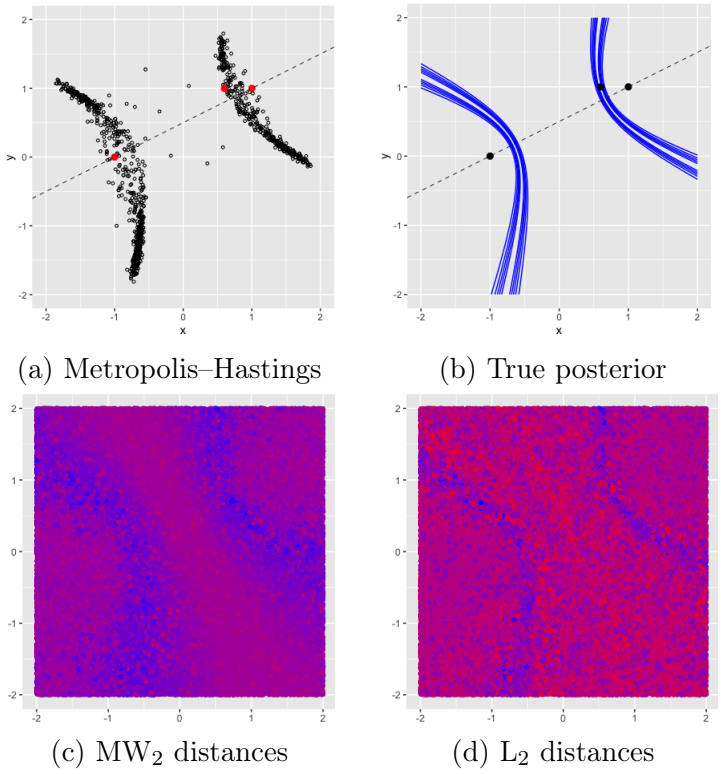
(a) GLliM-E-ABC (b) GLLiM-EV-ABC (c) GLLiM-MW2-ABC

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

Figure 4: Sound source localization. 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 third black point is the true sound source localization.

potential of multiple solutions or of the possibility to obtain very similar observations from different sets of parameters, which makes this setting appropriate for testing the ability of our GLLiM-D-ABC procedures to recover multimodal posterior distributions.

In the following experiments, all parameters are transformed to be in $[0,1]^4$, which amounts to keep b and c unchanged, divide $\overline{\theta}$ by 30 and operate the following change of variable for ω, $\gamma = 1 - \sqrt{1-\omega}$. This last transformation also has the advantage of avoiding the non-linearity of F_{Hapke}, when ω tends to 1. The experimental setting defines geometries at which the measurements are made, which in turn define F_{Hapke}. The number of geometries thus corresponds to the size d, of each observation. The measurement geometries used to define F_{Hapke} are borrowed from a real laboratory experiment presented below. The number of parameters is therefore 4 with $d = 10$ observed geometries. The size of the sets to learn the GLLiM model and generate ABC samples are both set to is $N = M = 10^5$. For



(a) Metropolis–Hastings (b) True posterior

(c) MW_2 distances (d) L_2 distances

Figure 5: Sound source localization. Sample from a Metropolis–Hastings algorithm (a) and contours of the true posterior distribution (b). Plots (c,d) show the MW_2 and L_2 distances before selection, high (resp. low) distances in red (resp. blue). Black points on the dotted line are the microphones positions. The third black point is the true sound source localization.

each couple $(\boldsymbol{\theta}, \mathbf{y})$ in the simulated data sets, the 4 parameters $(\boldsymbol{\theta})$ are simulated uniformly in $[0,1]^4$. The corresponding reflectance curves are generated as $\mathbf{y} = F_{\text{Hapke}}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a centered Gaussian variable with isotropic covariance $\sigma^2 \mathrm{I}_d$. In this section $\sigma = 0.05$. The GLLiM model is learned with $K = 40$. Previous studies reported in Kugler et al. (2020) showed that this value of K was satisfying.

6.3.1 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 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. The two signals are shown in Figure 6.

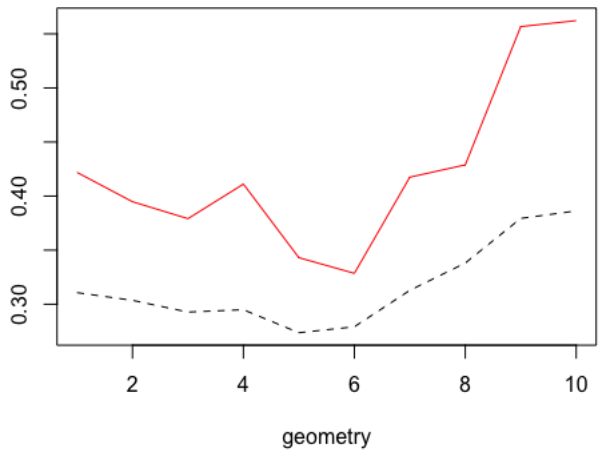


Figure 6: Synthetic signal (red line) used to illustrated the ABC procedures. The signal is chosen in the available data set as the one with the highest correlation to the real observed signal (black dashed line). This synthetic signal has been generated from the Hapke model applied to parameter values $(\omega, \bar{\theta}, b, c) = (0.68, 0.04, 0.23, 0.04)$ with an additional Gaussian noise with standard deviation of 0.05.

Figure 7 shows the marginal posteriors obtained for each parameter using the four ABC procedures and for different tolerance values ϵ chosen as the 0.05%, 0.1% and 1% quantile. 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 degres. The Hapke model is also such that ω and $\overline{\theta}$ values can interact to allow the reconstruction of a given spectrum. In Figure 7, 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 consistent with expert knowledge according to which b and $\overline{\theta}$ are more difficult to estimate than w and c.

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

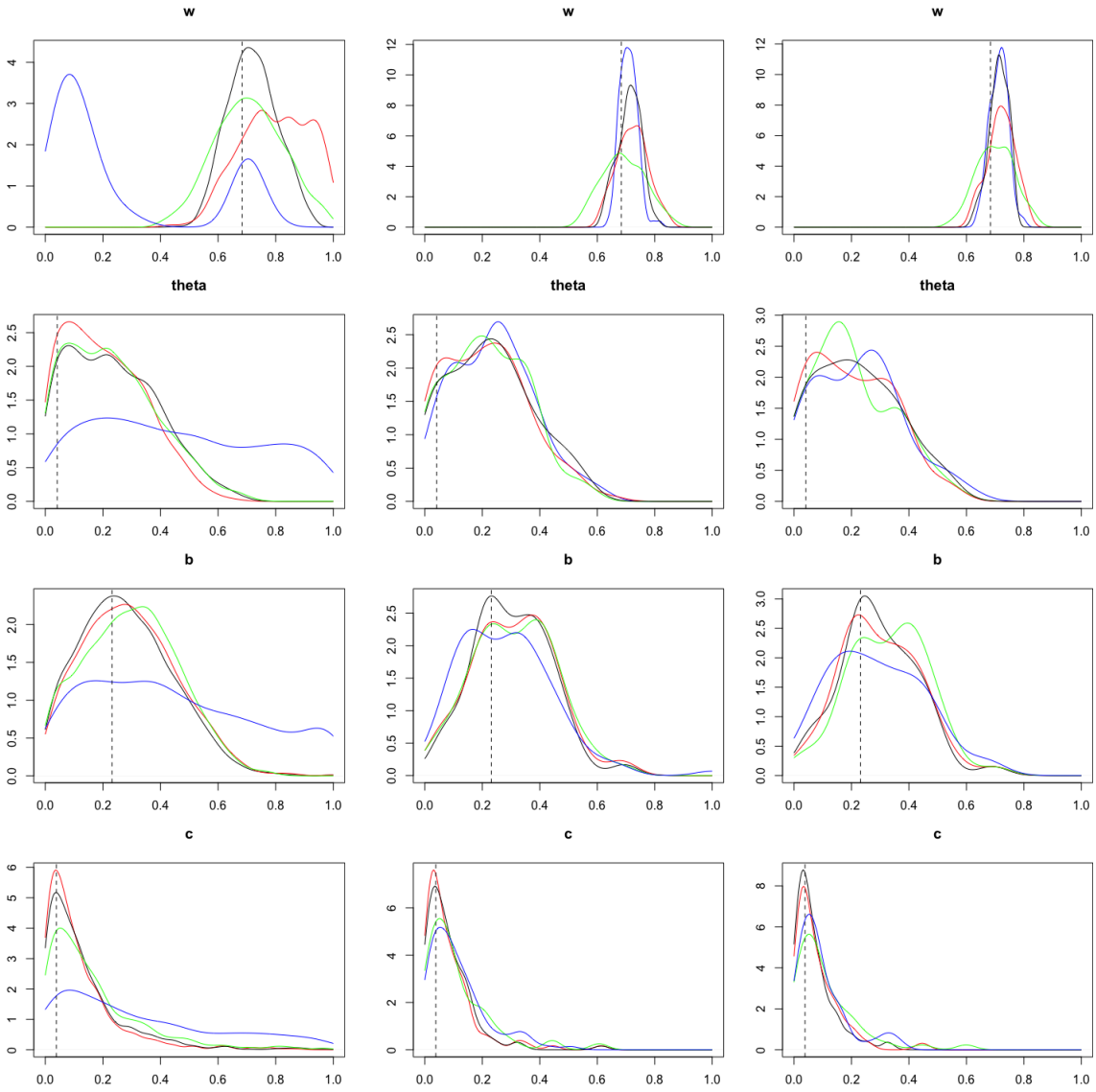
The GLLiM-MW2-ABC procedure shows a better ability to target the right parameter values, when compared to the GLLiM-L2-ABC procedure; see for instance the b posterior at 0.1%. It is interesting to vary ϵ to observe the behavior of the different methods. A lower ϵ can be used to check if one of the modes may vanish (*i.e.* with a more drastic thresholding) or is confirmed when the selection is more permissive. This is visible for the b parameter. The third column of Figure 7 shows that the GLLiM-MW2-ABC posterior has a clearer peak on the right parameter value, which is maintained with a larger variance in the first column. The GLLiM-L2-ABC procedure seems less robust to these variations and even degrades in performance when the thresholding is too permissive. The two procedures based on the expectations as summary statistics are also quite stable and have overall satisfying performance with globally less sharp posterior distributions.

6.3.2 Laboratory observations

Reflectance measurements made in the laboratory are also generally considered by experts (see *e.g.* Pilorget et al. 2016). As an illustration, we focus on one observation coming from an experiment involving a mineral called Nontronite (see also Kugler et al. 2020 for a recent description and study). The experiment consists of taking measures at 100 wavelengths in the spectral range 400–2800 nm. Each of these 100 measures is an observation to be inverted. We focus on one of them, at 2310 nm. This observation has been chosen from previous study (Kugler et al., 2020) as likely to exhibit multiple solutions. As before, the experimental setting defines geometries at which the measurements are made, which in turn define F_{Hapke}. The size d of each observation is $d = 10$ and the corresponding angles are such that the incidence and azimuth angles are fixed to $\theta_0 = 45$ and $\phi = 0$. This number d of geometries is quite typical of real observations for which the number of possible measurements during a planet flyover is limited.

Figure 8 provides the estimated posterior marginals for the Nontronite measurements, for each parameter. The shown distributions are obtained by setting ϵ to the 0.1% quantile of the computed distances.

From Figure 8, two solutions can be deduced. All parameters show unimodal posterior distributions except for $\overline{\theta}$, which exhibits two modes. Our results show that the multiplicity comes from the parameter $\overline{\theta}$. For such real data, no ground truth is available so that it is difficult to fully validate the estimations. However a simple inspection consists of checking the reconstructed signals. Figure 9 compares the inverted signal to the reconstructed signals obtained by applying the Hapke model to the two sets of estimated parameters, namely $(0.59, 0.15, 0.14, 0.06)$ and $(0.59, 0.42, 0.14, 0.06)$, which differ only in $\overline{\theta}$. The proximity of the reconstructions confirms the existence of multiple solutions, differing in $\overline{\theta}$, and thus the relevance of a multimodal posterior. However, one solution can be selected by selecting the set of parameters that provides the best reconstruction measured using the mean squared error (MSE). The set $(0.59, 0.42, 0.14, 0.06)$ is selected as its MSE is slightly lower (2.6×10^{-4} vs 3.3×10^{-4}). This is satisfactory, as the lower value of $\overline{\theta}$ in the other solution is less



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

Figure 7: 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, 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.

physically interpretable as mentioned earlier. Note that for the purpose of our comparison here and for simplicity, we used a uniform prior on $\boldsymbol{\theta}$ but for a more meaningful study in planetary science this information on the parameters plausible values could be incorporated directly to produce more informative priors.

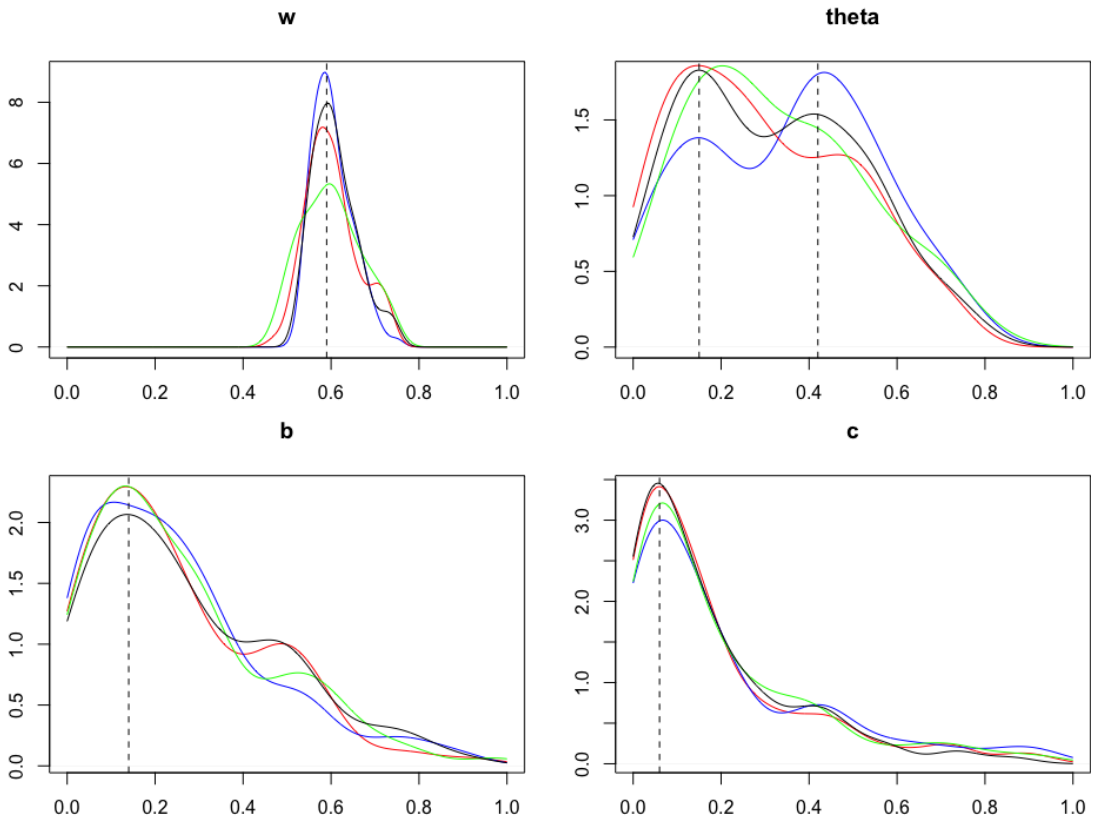


Figure 8: Real observation inversion using the Hapke model. Selected samples using four ABC methods, GLLiM-E-ABC expectations in red, semi-automatic ABC in green, GLLiM-L2-ABC in blue and GLLiM-MW2-ABC in black. The posterior margins for $\omega, \bar{\theta}, b$ and c are shown respectively. The threshold ϵ i set to the 0.1% quantile. 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 $\bar{\theta}$).

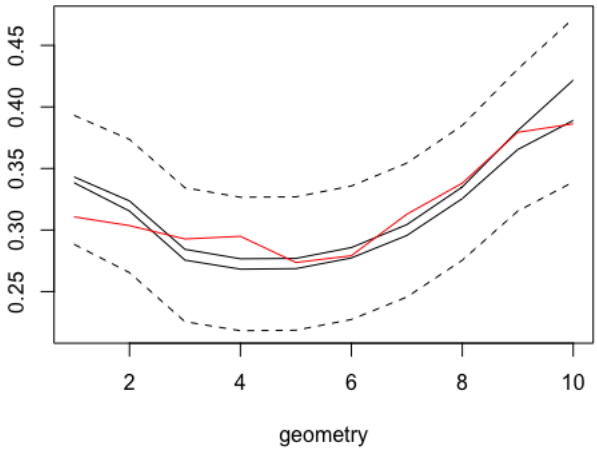


Figure 9: Signal reconstructions (black lines) obtained when applying the Hapke model to parameter values $(0.59, 0.15, 0.14, 0.06)$ and $(0.59, 0.42, 0.14, 0.06)$. The observed signal is shown in red. The dashed lines correspond to the addition/substraction of a standard deviation of 0.05 around the reconstructions.

7 Conclusion and perspectives

In this work, the issue of choosing summary statistics was revisited. We built on the seminal work of Fearnhead and Prangle (2012) and their semi-automatic ABC by replacing the approximate posterior expectations with functional statistics, namely approximations of the posterior distributions. These surrogate posterior distributions were obtained in a preliminary learning step, based on an inverse regression principle. This is original with respect to most standard regression procedures which usually provide only point-wise predictions, *i.e.* first order moments. So doing, we could not only compute approximate posterior moments of higher orders as summary statistics but more generally approximate full posterior distributions. More specifically, this learning step was based on the so-called GLLiM model, which provides surrogate posteriors in the parametric family of Gaussian mixtures. Preliminary experiments showed that although the posterior moments provided by GLLiM were not always leading to better results than that provided by semi-automatic ABC, the use of the full surrogate posteriors was always an improvement.

To handle distributions as summary statistics, our procedure required appropriate distances. We investigated L_2 and a Wassertein-based distance (MW_2), which are both tractable for mixtures of Gaussians. No significant differences between the two distances have been observed in our experiments but the MW_2 distance appeared to be more robust

in the sense of being less sensitive to small variations in the compared distributions. As illustrated in our remote sensing example, it may also allow for the ability to set the tolerance level at a higher value without overly degrading the quality of the posterior sample.

Among aspects that have not been thoroughly investigated in this work, we could refine the way to choose this tolerance level ϵ or combine GLLiM with more sophisticated ABC schemes than the simple rejection scheme.

In this current version, our proposal applies to ABC settings, where for a given parameter value, only one observation (that is possibly multi-dimensional) is available at a time. Such settings are of practical importance as they are typical of inverse problems, where many observations are measured but for different parameter values, due to experimental limitations or costs. In addition, even when more than one observation is available, it is common to use summary statistics. For instance, in their g-and-k distribution experiment, Fearnhead and Prangle (2012) consider for a true given parameter a sample of 10^4 observations but reduce it to 100 features to apply the regression step of their semi-automatic procedure. Similarly, Drovandi and Pettitt (2011) reduce their sample of 10^4 observations to a vector of 7 octiles. So doing their analyses imply the one observation scenario, that we consider. In contrast, methods using discrepancies (Bernton et al., 2019; Jiang et al., 2018) can handle samples directly and bypass the need for summary statistics. However, they require a relatively large number of generally *i.i.d.* observations for both the true and simulated parameters. For computational reasons, as for the semi-automatic procedure, the preliminary regression step in standard GLLiM is not adapted to the multiple observation case. Therefore, an important future direction is to extend this work to the case of *i.i.d.* samples. This requires the modification of the standard GLLiM procedure to maintain its approximation quality and computational efficiency. With this in mind, an important feature of GLLiM, not illustrated in this paper, is to allow the application of ABC procedures in high dimensional settings and to address the curse of dimensionality that is usually encountered in standard summary statistics based ABC. The rest of our proposal would then be easily adapted.

At last, in principle, any other method that is able to provide approximate surrogate posteriors could be used in place of GLLiM to produce the functional summaries. However, besides the family of mixture of experts models which are similar to GLLiM, to our knowledge, most regression techniques and typically neural networks focus on point-wise predictions.

Acknowledgements. FF would like to thank Guillaume Kon Kam King for an initial discussion on semi-automatic ABC which inspired this work, Benoit Kugler and Sylvain Douté for providing the simulations for the planetary science example and for helpful discussions on the Hapke model.

8 Proofs

8.1 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

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

8.2 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 $\epsilon < \epsilon(\alpha,\beta,\gamma), K \geq K(\alpha,\beta,\gamma), N \geq N(\alpha,\beta,\gamma)$:

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \geq \beta\right\}\right) > \alpha\right) \leq \gamma. \tag{29}$$

To prove (29), we first recall that we can rewrite $q_\epsilon^{K,N}$ as follows, for all $\boldsymbol{\theta} \in \Theta, \mathbf{y} \in \mathcal{Y}$,

$$\begin{aligned} q_\epsilon^{K,N}(\boldsymbol{\theta} \mid \mathbf{y}) &= \int_{\mathcal{Y}} K_\epsilon^{K,N}(\mathbf{z};\mathbf{y})\, \pi(\boldsymbol{\theta} \mid \mathbf{z})\, d\mathbf{z}, \\ K_\epsilon^{K,N}(\mathbf{z};\mathbf{y}) &= \frac{\mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}),p^{K,N}(\cdot|\mathbf{z}))\leq\epsilon\}}\pi(\mathbf{z})}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(p^{K,N}(\cdot|\mathbf{y}),p^{K,N}(\cdot|\mathbf{z}))\leq\epsilon\}}\pi(\mathbf{z})\, d\mathbf{z}}, \end{aligned} \tag{30}$$

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),

$$\left(\frac{1}{2} D_1(f,g)\right)^2 \leq D_H^2(f,g) \leq D_1(f,g). \tag{31}$$

Applying successively the right-hand-side of (31), 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

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

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

and using the left-hand-side of (31), we finally get that

$$D_{\mathrm{H}}^2\left(q_\epsilon^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) \leq 2D_{\mathrm{H}}\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right). \tag{32}$$

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 (32) can then be bounded by three terms using the triangle inequality for the Hellinger distance D_{H},

$$\begin{aligned} D_{\mathrm{H}}\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right) \leq\ & D_{\mathrm{H}}\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) + D_{\mathrm{H}}\left(\pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}\left(\cdot \mid \mathbf{y}\right)\right) \\ & + D_{\mathrm{H}}\left(p^{K,N}\left(\cdot \mid \mathbf{y}\right), \pi\left(\cdot \mid \mathbf{y}\right)\right). \end{aligned} \tag{33}$$

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 $\mathbf{z}_{0,\mathbf{y}}^{K,N} \in A_{0,\mathbf{y}}^{K,N} \subset \mathcal{Y}$,

$$\sup_{\boldsymbol{\theta}\in\Theta} \left|\pi\left(\boldsymbol{\theta} \mid \mathbf{z}\right) - \pi(\boldsymbol{\theta} \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right| \leq \alpha_1, \forall \mathbf{z} \in \mathcal{Y}, \left|\mathbf{z} - \mathbf{z}_{0,\mathbf{y}}^{K,N}\right| < \delta(\alpha_1). \tag{34}$$

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 (34), 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 (31) and (34), it follows for any ϵ such that $0 < \epsilon < \epsilon\left(\delta(\alpha_1)\right)$,

$$\begin{aligned} D_H^2\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) &\leq D_1\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \\ &\leq \sup_{\boldsymbol{\theta}\in\Theta} \left|\pi\left(\boldsymbol{\theta} \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right) - \pi(\boldsymbol{\theta} \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right| \lambda\left(\Theta\right) \\ &\leq \alpha_1 \lambda\left(\Theta\right). \end{aligned} \tag{35}$$

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 $0 < \epsilon < \epsilon\left(\alpha_1, \beta_1\right)$,

$$\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \geq \beta_1\right\}\right) \leq \alpha_1. \tag{36}$$

Then, since (36) 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

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(\pi\left(\cdot \mid \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}\right), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \geq \beta_1\right\}\right) > \alpha_1\right) = 0. \tag{37}$$

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 $D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$. 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

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

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

$$\begin{aligned}\int_{\mathcal{Y}} \mathrm{KL}\left(\pi(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{y})\right) d\lambda(\mathbf{y}) &\leq \frac{1}{a\lambda(\Theta)} \int_{\mathcal{Y}} \pi(\mathbf{y})\, \mathrm{KL}\left(\pi(\cdot \mid \mathbf{y}), p^{K,N}(\cdot \mid \mathbf{y})\right) d\lambda(\mathbf{y}) \\ &\leq \frac{1}{a\lambda(\Theta)}\, \mathrm{KL}\left(\pi, p^{K,N}\right), \end{aligned} \tag{38}$$

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 8.3.1), 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 $D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ converges to 0 in L_1 distance with respect to $\mathbf{y}$. Using Tao (2011, 1.5. Modes of convergence), $D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right)$ 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 $K(\alpha_2, \beta_2, \gamma_2) \in \mathbb{N}^*$, $N(\alpha_2, \beta_2, \gamma_2) \in \mathbb{N}^*$ such that for any $K \geq K(\alpha_2, \beta_2, \gamma_2)$, $N \geq N(\alpha_2, \beta_2, \gamma_2)$,

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y}, D_H^2\left(p^{K,N}(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{y})\right) \geq \beta_2\right\}\right) > \alpha_2\right) \leq \gamma_2. \tag{39}$$

To show that the same as (39) 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 8.3.2, guaranties first that the map $\mathbf{y} \mapsto \mathbf{z}_0^{K,N}(\mathbf{y}) = \mathbf{z}_{0,\mathbf{y}}^{K,N}$ 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 (39) implies that

$$\Pr\left(\lambda\left(\left\{\mathbf{y} \in \mathcal{Y} : D_H^2\left(p^{K,N}(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N}), \pi(\cdot \mid \mathbf{z}_{0,\mathbf{y}}^{K,N})\right) \geq \beta_2\right\}\right) > \alpha_2\right) \leq \gamma_2. \tag{40}$$

Finally, (29) can be deduced from (37), (39) and (40) 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 $N(\alpha, \beta, \gamma) = N(\alpha_2, \beta_2, \gamma_2)$.

8.3 Auxiliary results

8.3.1 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

$$\mathcal{C}_K = \mathrm{conv}_K(\mathcal{H}_{\mathcal{X}}) = \left\{f : f(\mathbf{x}) = \sum_{k=1}^{K} c_k g_{\boldsymbol{\varphi}_k}(\mathbf{x}),\, c \in \mathbb{S}^{K-1}, g_{\boldsymbol{\varphi}_k} \in \mathcal{H}_{\mathcal{X}}\right\} \tag{41}$$

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 $\mathrm{KL}(\pi, \mathcal{C}) = 0$. Using Lemma 1, it follows that for all $t > 0$, for all $K \in \mathbb{N}^*$, and for all $N \in \mathbb{N}^*$,

$$\Pr\left(\mathrm{KL}\left(\pi, p^{K,N}\right) \leq \frac{c_1}{K} + \frac{c_2}{\sqrt{N}} + \frac{c_3\sqrt{t}}{\sqrt{N}}\right) \geq 1 - \exp(-t)\,. \tag{42}$$

Choosing $t = N^{1/2}$, (42) becomes

$$1 - \Pr\left(\mathrm{KL}\left(\pi, p^{K,N}\right) \leq \frac{c_1}{K} + \frac{c_2}{\sqrt{N}} + \frac{c_3}{N^{1/4}}\right) \leq \exp\left(-N^{1/2}\right). \tag{43}$$

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 (43) 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 proof that $\mathrm{KL}\left(\pi, p^{K,N}\right) \to 0$, in probability as $K \to \infty, N \to \infty$.

8.3.2 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

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

Then if $\mathcal{Z}$ is compact 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,*

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

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

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* $B^{K,N}_{0,\mathbf{y}} = \underset{\mathbf{z} \in A^{K,N}_{0,\mathbf{y}}}{\arg\max} D_1(\pi(\cdot \mid \mathbf{z}), \pi(\cdot \mid \mathbf{y}))$.

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

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

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

References

Aliprantis, C. D. and Border, K. C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media. (Cited on pages 39, 40, and 41.)

Arridge, S., Maass, P., Öktem, O., and Schönlieb, C.-B. (2019). Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174. (Cited on page 6.)

Beal, M. J., Jojic, N., and Attias, H. (2003). A graphical model for audiovisual object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(7):828–836. (Cited on page 22.)

Bernard-Michel, C., Douté, S., Fauvel, M., Gardes, L., and Girard, S. (2009). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research: Planets*, 114(E6). (Cited on page 3.)

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81:235–269. (Cited on pages 3, 5, 12, 13, 16, 32, and 33.)

Blum, M. G. B., Nunes, M. A., Prangle, D., and Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208. (Cited on page 3.)

Buchholz, A. and Chopin, N. (2019). Improving Approximate Bayesian Computation via Quasi-Monte Carlo. *Journal of Computational and Graphical Statistics*, 28(1):205–219. (Cited on page 3.)

Chen, Y., Georgiou, T. T., and Tannenbaum, A. (2019). Optimal Transport for Gaussian Mixture Models. *IEEE Access*, 7:6269–6278. (Cited on pages 4, 5, and 8.)

Cook, R. D. and Forzani, L. (2019). Partial least squares prediction in high-dimensional regression. *The Annals of Statistics*, 47(2):884–908. (Cited on page 6.)

Csillery, K., Francois, O., and Blum, M. G. B. (2012). abc: an R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*. (Cited on page 16.)

Del Moral, P., Doucet, A., and Jasra, A. (2012). An Adaptive Sequential Monte Carlo Method for Approximate Bayesian Computation. *Statistics and Computing*, 22(5):1009–1020. (Cited on page 3.)

Deleforge, A., Forbes, F., Ba, S., and Horaud, R. (2015a). Hyper-Spectral Image Analysis with Partially-Latent Regression and Spatial Markov Dependencies. *IEEE Journal of Selected Topics in Signal Processing*, 9(6):1037–1048. (Cited on page 3.)

Deleforge, A., Forbes, F., and Horaud, R. (2015b). High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables. *Statistics and Computing*, 25(5):893–911. (Cited on pages 4, 6, and 8.)

Delon, J. and Desolneux, A. (2020). A Wasserstein-type distance in the space of Gaussian Mixture Models. *SIAM Journal on Imaging Sciences*. (Cited on pages 4, 5, 8, and 9.)

Drovandi, C. C. and Pettitt, A. N. (2011). Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics and Data Analysis*, 55:2541–2556. (Cited on page 32.)

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474. (Cited on pages 3, 4, 7, 16, 31, and 32.)

Fernando, J., Schmidt, F., and Douté, S. (2016). Martian surface microtexture from orbital CRISM multi-angular observations: A new perspective for the characterization of the geological processes. *Planetary and Space Science*, 128:30–51. (Cited on page 24.)

Geyer, C. J. and Jonhson, L. T. (2020). mcmc: Markov chain Monte Carlo. https://cran.r-project.org/web/packages/mcmc/. (Cited on page 23.)

Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2018). Likelihood-free inference via classification. *Statistics and Computing*, 28:411–425. (Cited on pages 3, 5, and 16.)

Hospedales, T. M. and Vijayakumar, S. (2008). Structure inference for Bayesian multisensory scene understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(12):2140–2157. (Cited on page 22.)

Hovorka, R., Canonico, V., Chassin, L. J., Haueter, U., Massi-Benedetti, M., Federici, M. O., Pieber, T. R., Schaller, H. C., Schaupp, L., Vering, T., and Wilinska, M. E. (2004). Nonlinear model predictive control of glucose concentration in subjects with type 1 diabetes. *Physiological Measurement*, 25(4):905–920. (Cited on page 2.)

Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local Statistical Modeling via a Cluster-Weighted Approach with Elliptical Distributions. *Journal of classification*, 29(3):363–401. (Cited on page 6.)

Jiang, B., Wu, T.-Y., C., Z., and Wong, W. (2017). Learning summary statistics for Approximate Bayesian Computation via Deep Neural Network. *Statistica Sinica*, pages 1595–1618. (Cited on pages 3, 4, 7, and 16.)

Jiang, B., Wu, T.-Y., and Wong, W. H. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*. (Cited on pages 5, 18, and 32.)

Kristan, M., Leonardis, A., and Skočaj, D. (2011). Multivariate online kernel density estimation with Gaussian kernels. *Pattern Recognition*, 44(10-11):2630–2642. (Cited on page 5.)

Kugler, B., Forbes, F., and Douté, S. (2020). Fast Bayesian Inversion for high dimensional inverse problems. *https://hal.archives-ouvertes.fr/hal-02908364*. (Cited on pages 24, 26, and 28.)

Labarre, S. (2017). *Caractérisation et modélisation de la rugosité multi-échelle des surfaces naturelles par télédétection dans le domaine solaire*. PhD thesis, Physique Univers Sorbonne Paris Cité. Supervised by C. Ferrari and S. Jacquemoud. (Cited on page 24.)

Lemasson, B., Pannetier, N., Coquery, N., Boisserand, L. S. B., Collomb, N., Schuff, N., Moseley, M., Zaharchuk, G., Barbier, E. L., and Christen, T. (2016). MR Vascular Fingerprinting in Stroke and Brain Tumors Models. *Scientific Reports*, 6:37071. (Cited on page 3.)

Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *J. Amer. Stat. Assoc.*, 86(414):316–327. (Cited on page 6.)

Ma, D., Gulani, V., Seiberlich, N., Liu, K., Sunshine, J. L., Duerk, J. L., and Griswold, M. A. (2013). Magnetic Resonance Fingerprinting. *Nature*, 495(7440):187–192. (Cited on page 3.)

Makarov, B. and Podkorytov, A. (2013). *Real analysis: measures, integrals and applications*. Springer Science & Business Media. (Cited on pages 34 and 37.)

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computation methods. *Statistics and Computing*, 22:1167–1180. (Cited on pages 18 and 21.)

Mesejo, P., Saillet, S., David, O., Bénar, C., Warnking, J. M., and Forbes, F. (2016). A differential evolution-based approach for fitting a nonlinear biophysical model to fMRI BOLD data. *IEEE Journal of Selected Topics in Signal Processing*, 10(2):416–427. (Cited on page 2.)

Muandet, K., Fukumizu, K., Dinuzzo, F., and Scholkopf, B. (2012). Learning from distributions via support measure machines. In *Advances in neural information processing systems*, pages 10–18. (Cited on page 5.)

Murchie, S. L., Seelos, F. P., Hash, C. D., Humm, D. C., Malaret, E., McGovern, J. A., Choo, T. H., Seelos, K. D., Buczkowski, D. L., Morgan, M. F., Barnouin-Jha, O. S.,

Nair, H., Taylor, H. W., Patterson, G. W., Harvel, C. A., Mustard, J. F., Arvidson, R. E., McGuire, P., Smith, M. D., Wolff, M. J., Titus, T. N., Bibring, J.-P., and Poulet, F. (2009). Compact Reconnaissance Imaging Spectrometer for Mars investigation and data set from the Mars Reconnaissance Orbiter's primary science phase. *Journal of Geophysical Research: Planets*, 114(E2):E00D07. (Cited on page 24.)

Nataraj, G., Nielsen, J.-F., Scott, C., and Fessler, J. A. (2018). Dictionary-Free MRI PERK: Parameter Estimation via Regression with Kernels. *IEEE Trans. Med. Imaging*, 37(9):2103–2114. (Cited on page 6.)

Nguyen, H. D., Arbel, J., Lü, H., and Forbes, F. (2020). Approximate Bayesian Computation Via the Energy Statistic. *IEEE Access*, 8:131683–131698. (Cited on pages 5, 8, and 18.)

Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2019). Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. *Neurocomputing*. (Cited on page 6.)

Nunes, M. A. and Prangle, D. (2015). abctools: An R package for tuning Approximate Bayesian Computation analyses. https://cran.r-project.org/web/packages/abctools/. (Cited on page 16.)

Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-ABC: approximate Bayesian computation with kernel embeddings. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*. (Cited on page 5.)

Perthame, E., Forbes, F., Deleforge, A., Devijver, E., and Gallopin, M. (2017). xLLiM: An R package for High Dimensional Locally-Linear Mapping. https://cran.r-project.org/web/packages/xLLiM/. (Cited on page 17.)

Pilorget, C., Fernando, J., Ehlmann, B. L., Schmidt, F., and Hiroi, T. (2016). Wavelength dependence of scattering properties in the VIS–NIR and links with grain-scale physical and compositional properties. *Icarus*, 267:296–314. (Cited on page 28.)

Prangle, D., Everitt, R. G., and Kypraios, T. (2018). A rare event approach to high-dimensional approximate Bayesian computation. *Statistics and Computing*, 28:819–834. (Cited on page 11.)

Rakhlin, A., Panchenko, D., and Mukherjee, S. (2005). Risk bounds for mixture density estimation. *ESAIM: Probability and Statistics*, 9:220–229. (Cited on pages 2, 15, 36, 37, and 38.)

Rubio, F. and Johansen, A. M. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, 7:1632–1654. (Cited on page 12.)

Schmidt, F. and Fernando, J. (2015). Realistic uncertainties on Hapke model parameters from photometric measurements. *Icarus*, 260:73–93 (IF 2,84). (Cited on pages 3 and 24.)

Sisson, S. A., Fan, Y., and Beaumont, M. A., editors (2019). *Handbook of Approximate Bayesian Computation*. CRC Press, Boca Raton. (Cited on page 2.)

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Scholkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561. (Cited on page 5.)

Tao, T. (2011). *An introduction to measure theory*. American Mathematical Society Providence, RI. (Cited on pages 36 and 37.)

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media. (Cited on page 36.)

Wang, D. and Brown, G. J. (2006). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Wiley-IEEE Press. (Cited on page 22.)

Wang, F., Syeda-Mahmood, T., Vemuri, B. C., Beymer, D., and Rangarajan, A. (2009). Closed-form Jensen-Renyi divergence for mixture of Gaussians and applications to group-wise shape registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 648–655. Springer. (Cited on page 5.)

Wiqvist, S., Mattei, P.-A., Picchini, U., and Frellsen, J. (2019). Partially exchangeable networks and architectures for learning summary statistics in approximate Bayesian computation. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6798–6807, Long Beach, California, USA. (Cited on pages 3 and 7.)

Zeevi, A. J. and Meir, R. (1997). Density estimation through convex combinations of densities: approximation and estimation bounds. *Neural Networks*, 10(1):99–109. (Cited on page 34.)