



HAL
open science

Approximate Bayesian computation with surrogate posteriors

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

► **To cite this version:**

Florence Forbes, Hien Duy Nguyen, Trung Tin Nguyen, Julyan Arbel. Approximate Bayesian computation with surrogate posteriors. 2021. hal-03139256v4

HAL Id: hal-03139256

<https://hal.science/hal-03139256v4>

Preprint submitted on 6 May 2022 (v4), last revised 25 Sep 2022 (v5)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Approximate Bayesian computation with surrogate posteriors

Florence Forbes^{1*}, Hien Duy Nguyen², TrungTin Nguyen³
and Julyan Arbel¹

¹*Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Inria
Grenoble Rhone-Alpes, 655 av. de l'Europe, 38335 Montbonnot,
France.

²School of Engineering and Mathematical Sciences, La Trobe
University, Bundoora, Victoria, Australia.

³Normandie Univ, UNICAEN, CNRS, LMNO, Caen, 14000,
France.

*Corresponding author(s). E-mail(s): florence.forbes@inria.fr;
Contributing authors: H.Nguyen5@latrobe.edu.au;
trung-tin.nguyen@unicaen.fr; julyan.arbel@inria.fr;

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 thus their choice can affect the quality of approximations. The choice between discrepancies is an active research topic, which has mainly considered data discrepancies requiring samples of observations or distances between summary statistics. 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 a functional summary statistics

001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046

047 setting and can also be used as an alternative to sample-based ap-
 048 proaches. The resulting ABC quasi-posterior distribution is shown to
 049 converge to the true one, under standard conditions. Performance is il-
 050 lustrated on both synthetic and real data sets, where it is shown that
 051 our approach is particularly useful when the posterior is multimodal.

052 **Keywords:** Approximate Bayesian computation, summary statistics,
 053 surrogate models, Gaussian mixtures, Wasserstein distance, multimodal
 054 posterior distributions.

055

056

057

058 1 Introduction

059

060 Approximate Bayesian computation (ABC) (see, *e.g.*, [Sisson et al. 2019](#)) ap-
 061 pears as a natural candidate for addressing problems, where there is a lack of
 062 availability or tractability of the likelihood. Such cases occur when the direct
 063 model or data generating process is not available analytically, but is available
 064 as a simulation procedure; *e.g.*, when the data generating process is charac-
 065 terized as a series of ordinary differential equations, as in [Mesejo et al. \(2016\)](#)
 066 or [Hovorka et al. \(2004\)](#). In addition, typical features or constraints that can
 067 occur in practice are that: 1) the observations \mathbf{y} are high-dimensional, because
 068 they represent signals in time or are spectral, as in [Schmidt and Fernando](#)
 069 [\(2015\)](#); [Bernard-Michel et al. \(2009\)](#); [Ma et al. \(2013\)](#); and 2) the parameter θ ,
 070 to be estimated, is itself multi-dimensional with correlated dimensions so that
 071 independently predicting its components is sub-optimal; *e.g.*, when there are
 072 known constraints such as when the parameter elements are concentrations or
 073 probabilities that sum to one ([Deleforge et al., 2015a](#); [Lemasson et al., 2016](#);
 074 [Bernard-Michel et al., 2009](#)).

075 The fundamental idea of ABC is to generate parameter proposals θ in a
 076 parameter space Θ using a prior distribution $\pi(\theta)$ and accept a proposal if
 077 the simulated data \mathbf{z} for that proposal is similar to the observed data \mathbf{y} , both
 078 in an observation space \mathcal{Y} . This similarity is usually measured using a dis-
 079 tance or discriminative measure D and a simulated sample \mathbf{z} is retained if
 080 $D(\mathbf{z}, \mathbf{y})$ is smaller than a given threshold ϵ . In this simple form, the procedure
 081 is generally referred to as rejection ABC. Other variants are possible and of-
 082 ten recommended, for instance using MCMC or sequential procedures (*e.g.*,
 083 [Del Moral et al., 2012](#); [Buchholz and Chopin, 2019](#)). We will focus on the re-
 084 jection version for the purpose of this paper as all developments in this setting
 085 can be easily adapted to more sophisticated variants. Our analysis focuses on
 086 the convergence of the ABC quasi-posterior, as ϵ vanishes, which has only been
 087 studied in the context of rejection algorithms, to the best of our knowledge.
 088 However, we also illustrate the use of sequential Monte Carlo (SMC)-ABC in
 089 our numerical experiments.

090 In the case of a rejection algorithm, selected samples are drawn from the
 091 so-called ABC quasi-posterior, which is an approximation to the true posterior

092

$\pi(\boldsymbol{\theta} \mid \mathbf{y})$. 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

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

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 overload D to also 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 \(2012\)](#), the authors suggested to use a linear regression model to approximate $\mathbb{E}[\boldsymbol{\theta} \mid \mathbf{y}]$. This is very efficient in a number of settings. However, it is easy to construct examples, as illustrated in [Jiang et al. \(2017\)](#), [Wiqvist et al. \(2019\)](#) and [Akesson et al. \(2021\)](#), for which the approximation requires a richer approximation class. Still focusing on posterior means as summary statistics, the cited works 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. In addition, as shown by [Chen et al. \(2021\)](#), the choice of s as the posterior mean may lead to loss of information about the posterior distribution. [Chen et al. \(2021\)](#) propose instead to target a near sufficient statistics using a mutual information criterion.

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

139 (Deleforge et al., 2015b), that we recall in Section 3, appears as a good can-
140 didate, with properties that balance between the computationally expensive
141 neural networks and the simple standard regression techniques. In contrast
142 to most regression methods that provide only pointwise predictions, GLLiM
143 provides, at low cost, a parametric estimation of the full true posterior distribu-
144 tion. Using a learning set of parameters and observations pairs, GLLiM learns
145 a family of finite Gaussian mixtures whose parameters depend analytically on
146 the observation to be inverted. For any observed data, the true posterior can
147 be approximated as a Gaussian mixture, whose moments are easily computed
148 and turned into summary statistics for subsequent ABC sample selection.

149 Our second contribution is to propose to compare directly the full surrogate
150 posterior distributions provided by GLLiM, without reducing them to their
151 moments. So doing, we use a notion of functional summary statistics, which
152 also requires a different notion of the usual distances or discrepancy measures
153 to compare them. Recent developments in optimal transport-based distances
154 designed for Gaussian mixtures (Delon and Desolneux, 2020; Chen et al., 2019)
155 match perfectly this need via the so-called Mixture-Wasserstein distance as
156 referred to by Delon and Desolneux (2020), and denoted throughout the text
157 as MW_2 . There exist other distances between mixtures that are tractable, and
158 among them, the L_2 distance is also considered in this work.

159 A remarkable feature of our approach is that it can be equally applied to
160 settings where a sample of *i.i.d.* observations is available (*e.g.* Bernton et al.
161 (2019); Nguyen et al. (2020a)) and to settings where a single observation is
162 available, as a vector of measures, a time series realization or a data set reduced
163 to a vector of summary statistics (*e.g.* Fearnhead and Prangle (2012); Drovandi
164 and Pettitt (2011)).

165 The novelty of our approach and its comparison with existing work is em-
166 phasized in Section 2. The GLLiM output is briefly described in Section 3.
167 A first exploitation of GLLiM combined with the semi-automatic ABC prin-
168 ciple is presented in Section 4.1. Our extension, using functional summary
169 statistics, is then described in Section 4.2. The approach's theoretical proper-
170 ties are investigated in Section 5 and the practical performance is illustrated
171 in Section 6, both on synthetic and real data. Then, Section 7 concludes
172 the paper and discusses perspectives. Detailed proofs and additional illustra-
173 tions are shown in a supplementary material file. The code can be found at
174 <https://github.com/Trung-TinNguyenDS/GLLiM-ABC>.

175

176 2 Related work

177

178 As an alternative to semi-automatic ABC, in the works of Nguyen et al.
179 (2020a); Jiang et al. (2018); Bernton et al. (2019); Park et al. (2016); Gutmann
180 et al. (2018), the difficulties associated with finding efficient summary statistics
181 were bypassed by adopting, respectively, the Energy Distance, a Kullback-
182 Leibler divergence estimator, the Wasserstein distance, the Maximum Mean
183

184

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. \(2019\)](#) and [Gutmann et al. \(2018\)](#) who proposed solutions to also handle time series. These methods require sufficiently large samples and cannot be applied if the sample related to the parameter to be recovered is too small. This is a major difference with the approach we propose, which can be applied in both cases. We refer to these two cases as the *one observation* and *i.i.d. observations* settings. In the *one observation* case, the observed data restricts to a single observation \mathbf{y} of dimension d assumed to be generated from a true parameter $\boldsymbol{\theta}$ of dimension ℓ . This case is commonly encountered in inverse problems where it may be impossible to gather repeated observations from the same parameter values due to technological reasons. Typically, in remote sensing applications, satellites are limited to only a few degrees of freedom when observing a given site in constant conditions. This is also the case when the observation is a time series or when a sample of observations is reduced to a single vector of summary statistics. In the multiple *i.i.d.* observations case, the observed data is made of a sample of R *i.i.d.* realizations $\{\mathbf{y}^1, \dots, \mathbf{y}^R\}$ coming from the same true $\boldsymbol{\theta}$. The previous case is trivially recovered when $R = 1$.

ABC procedures using a regression step, as introduced by [Fearnhead and Prangle \(2012\)](#), are adapted to one observation settings. They cannot be applied on large (*e.g.* $R = 10^4$) numbers of covariates and require that samples, observed and simulated, are first reduced to a smaller number of statistics, *e.g.* 100. In contrast, discrepancy-based approaches compare empirical distributions constructed from the samples and require a relatively large R .

Our method is not limited to either one of these cases because we do not compare samples from distributions, but directly the distributions through their surrogates using distances between distributions. We can use the same Wasserstein, Kullback–Leibler divergence, *etc.*, but in their *population* versions rather than in their empirical versions. A Wasserstein-based distance can be computed between mixtures of Gaussians, thanks to the recent work of [Delon and Desolneux \(2020\)](#) and [Chen et al. \(2019\)](#). 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 unscented transform in order to compute the Hellinger distance between two Gaussian mixtures, although it is unclear what the complexity of this algorithm is.

To emphasize the difference to more standard summaries, we refer to our surrogate posteriors as functional summary statistics. The term has already been used by [Soubeyrand et al. \(2013\)](#) in the ABC context in their attempts to characterize spatial structures using statistics that are functions (*e.g.* correlograms or variograms). They do not address the issue of choosing the summary statistics. Given such functional statistics whose nature may change for each

185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230

231 considered model, their goal is to optimize the distances to compare them. In
 232 our proposal, the functional statistics are probability distributions. They arise
 233 as a way to bypass the summary statistics choice, but in this work, we make
 234 use of existing metrics to compare them, without optimization. We make note
 235 that the nomenclature: *functional summary statistics*, has also been used in a
 236 similar way by [Rodrigues et al. \(2016\)](#), where ABC is used to estimate func-
 237 tional, infinite dimensional, objects. Such objects are compared via simulated
 238 samples which are themselves summarized using kernel density estimators.
 239 These kernel densities are seen as functional summaries but they are not di-
 240 rectly related to a surrogate of the posterior distribution. The approach by
 241 [Rodrigues et al. \(2016\)](#) is closer to data discrepancy-based methods such as in
 242 [Nguyen et al. \(2020a\)](#); [Jiang et al. \(2018\)](#); [Bernton et al. \(2019\)](#); [Park et al.](#)
 243 [\(2016\)](#); [Gutmann et al. \(2018\)](#), that all require samples to compute meaningful
 244 nonparametric summaries, *e.g.* histograms.

245

246 **3 Parametric posterior approximation with** 247 **Gaussian mixtures**

248

249 A learning set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ is built from the joint distri-
 250 bution that results from the prior $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ and the likelihood $f_{\boldsymbol{\theta}}$, where
 251 $[N] = \{1, \dots, N\}$. More specifically, each pair $(\boldsymbol{\theta}_n, \mathbf{y}_n)$ in \mathcal{D}_N is obtained by
 252 simulating $\boldsymbol{\theta}_n$ from the prior $\pi(\boldsymbol{\theta})$ and \mathbf{y}_n from the likelihood $f_{\boldsymbol{\theta}_n}(\mathbf{y})$. The idea
 253 is to capture the relationship between $\boldsymbol{\theta}$ and \mathbf{y} with a joint probabilistic model
 254 for which computing conditional distributions and moments is straightforward.
 255 For the choice of the model to fit to \mathcal{D}_N , we propose to use the so-called Gaus-
 256 sian Locally Linear Mapping (GLLiM) model ([Deleforge et al., 2015b](#)) for its
 257 ability to capture non-linear relationships in a tractable manner, based on
 258 flexible mixtures of Gaussian distributions. GLLiM can be considered within
 259 the class of inverse regression approaches, such as sliced inverse regression ([Li,](#)
 260 [1991](#)), partial least squares ([Cook and Forzani, 2019](#)), mixtures of regressions
 261 approaches of different variants, *e.g.* mixtures of experts ([Nguyen et al., 2019](#)),
 262 cluster weighted models ([Ingrassia et al., 2012](#)), and kernel methods ([Nataraj](#)
 263 [et al., 2018](#)). In contrast to most deep learning approaches (see [Arridge et al.](#)
 264 [2019](#), for a survey), GLLiM provides for each observed \mathbf{y} , a full posterior proba-
 265 bility distribution within a family of parametric models $\{p_G(\boldsymbol{\theta} | \mathbf{y}; \boldsymbol{\phi}), \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$.
 266 Notable exceptions include mixture density networks (MDN, [Bishop \(1994\)](#)),
 267 which provide full posterior distributions as mixtures of Gaussians, and more
 268 generally normalizing flows ([Dinh et al., 2015](#)). These approaches could be
 269 considered instead of GLLiM with some adaptation (see the discussion in the
 270 conclusion, Section 7). To model non-linear relationships, GLLiM uses a mix-
 271 ture of K linear models. More specifically, the expression of $p_G(\boldsymbol{\theta} | \mathbf{y}; \boldsymbol{\phi})$ is
 272 analytical and available for all \mathbf{y} with $\boldsymbol{\phi}$ being independent of \mathbf{y} :

273

274

275

276

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

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian pdf with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $\eta_k(\mathbf{y}) = \pi_k \mathcal{N}(\mathbf{y}; \mathbf{c}_k, \boldsymbol{\Gamma}_k) / \sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{y}; \mathbf{c}_j, \boldsymbol{\Gamma}_j)$. This distribution involves parameters: $\boldsymbol{\phi} = \{\pi_k, \mathbf{c}_k, \boldsymbol{\Gamma}_k, \mathbf{A}_k, \mathbf{b}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$. One interesting property of this model is that the mixture setting provides guarantees that, when choosing K large enough, it is possible to approximate any reasonable relationship (Nguyen et al., 2019, 2020b,c, 2021a). The parameter $\boldsymbol{\phi}$ can be estimated by fitting a GLLiM model to \mathcal{D}_N using an Expectation-Maximization (EM) algorithm. Details are provided in supplementary material and in Deleforge et al. (2015b). In terms of learning, the GLLiM model has a $\mathcal{O}(Kd\ell)$ number of parameters to be estimated. The exact number of parameters depends on the variant learned. A reasonable size for the training data set then depends mainly on the number of parameters.

Fitting a GLLiM model to \mathcal{D}_N therefore results in a set of parametric distributions $\{p_G(\boldsymbol{\theta} | \mathbf{y}; \boldsymbol{\phi}_{K,N}^*), \mathbf{y} \in \mathcal{Y}\}$, 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} . When required, it is straightforward to compute the expectation and covariance matrix of $p_G(\boldsymbol{\theta} | \mathbf{y}; \boldsymbol{\phi}_{K,N}^*)$ in (2):

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

$$\begin{aligned} \text{Var}_G[\boldsymbol{\theta} | \mathbf{y}; \boldsymbol{\phi}_{K,N}^*] &= \sum_{k=1}^K \eta_k^*(\mathbf{y}) [\boldsymbol{\Sigma}_k^* + (\mathbf{A}_k^* \mathbf{y} + \mathbf{b}_k^*)(\mathbf{A}_k^* \mathbf{y} + \mathbf{b}_k^*)^\top] \\ &\quad - \mathbb{E}_G[\boldsymbol{\theta} | \mathbf{y}; \boldsymbol{\phi}_{K,N}^*] \mathbb{E}_G[\boldsymbol{\theta} | \mathbf{y}; \boldsymbol{\phi}_{K,N}^*]^\top. \end{aligned} \quad (4)$$

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

When R *i.i.d.* d -dimensional observations are available for each parameter value, they can be stacked into a single large vector. However, as noted by Fearnhead and Prangle (2012) and Jiang et al. (2017), the resulting number of covariates, of dimension at least $d \times R$, may become too large. Even if this is computationally doable with the standard GLLiM procedure, it is likely to be sub-optimal as it ignores the *i.i.d.* nature of the data. To handle this case, we therefore propose an adaptation of the EM algorithm of Deleforge et al. (2015b). This adaptation, referred to as GLLiM-iid, is detailed in the supplementary material Section S1 and illustrated in the first three examples of Section 6. It is shown by Deleforge et al. (2015b) that constraints on the model parameterization can be assumed without oversimplifying mixture (2). These constraints concern the covariance matrices used in the mixture modeling of the likelihood (or the direct model) and are not directly visible on the $\boldsymbol{\Sigma}_k$'s

323 which remain full in general. In addition to model the *i.i.d.* case, the adaptation
 324 we propose adds to the existing constraints, isotropic or diagonal matrices, the
 325 possibility to assume block diagonal structures.

326 In addition to choosing the covariance structure, GLLiM requires the choice
 327 of K the number of Gaussian components. Recent results by [Nguyen et al.](#)
 328 (2019, 2021a) justify a somewhat arbitrary choice of K , provided that it is suf-
 329 ficiently large. Intuitively, highly non-linear likelihoods may require a greater
 330 K . Previous studies have shown that the exact value of K was not critical (*e.g.*
 331 [Boux et al. \(2021\)](#)). This is also what we observed in our experiments com-
 332 paring different values of K (see Sections 6.3 and 6.4). A larger K provides
 333 generally better predictions but marginally so above a certain value. Never-
 334 theless, statistical selection procedures exist to choose K in a principled way.
 335 For instance in the paper introducing GLLiM, [Deleforge et al. \(2015b\)](#), the
 336 Bayesian Information Criterion (BIC) was used to select K and shows good re-
 337 sults. The authors in [Nguyen et al. \(2021b\)](#) also illustrate that non asymptotic
 338 approaches such as the slope heuristic, supported by non-asymptotic oracle
 339 inequalities, can also work well for GLLiM on synthetic and real datasets. Al-
 340 ternatively to standard information criteria, a Bayesian nonparametric version
 341 of GLLiM could be implemented not to commit to an arbitrary K value. In
 342 practical inverse problems, the choice of K can also be guided by the qual-
 343 ity of the learned direct model, which only requires a learning data set to be
 344 evaluated.

345

346 4 Extended semi-automatic ABC

347

348 Semi-automatic ABC refers to an approach introduced in [Fearnhead and Pran-](#)
 349 [gle \(2012\)](#), which has since then led to various attempts and improvements,
 350 see *e.g.* [Jiang et al. \(2017\)](#), [Wiqvist et al. \(2019\)](#) and [Akesson et al. \(2021\)](#),
 351 without dramatic deviation from the original ideas.

352

353 4.1 Extension to extra summary vectors

354

355 A natural idea is to use the approximate posterior expectation provided by
 356 GLLiM in (3) as the summary statistic s of data \mathbf{y} , $s(\mathbf{y}) = \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}_{K,N}^*]$.
 357 It provides a first attempt to combine GLLiM and ABC procedures and has
 358 the advantage over neural networks of being easier to estimate without the
 359 need for complex hyperparameter tuning. GLLiM requires only the setting of
 360 an integer parameter K , while neural networks require the choice of a full
 361 architecture, number of layers, number of nodes per layer, etc.

362

363 However, one advantage of GLLiM over most regression methods is not to
 364 reduce to pointwise predictions and to provide full posteriors as output. The
 365 posteriors can then be used to provide other posterior moments as summary
 366 statistics. The same standard ABC procedure as before can be applied but now
 367 with $s_1(\mathbf{y}) = \mathbb{E}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}_{K,N}^*]$ and $s_2(\mathbf{y}) = \text{Var}_G[\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}_{K,N}^*]$, as given by (4).

367

368 As illustrated in Section 6, it is easy to construct examples where the
 posterior expectations, even when well-approximated, do not perform well as

summary statistics. See also Proposition 2 in [Chen et al. \(2021\)](#) for a more theoretical justification. 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 moments.

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 approximations 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, as can the Hellinger distance or any other divergence. A natural choice is the Kullback–Leibler divergence, but computing it between mixtures is not straightforward. Computing the Energy statistic (*e.g.*, [Nguyen et al., 2020a](#)) appears at first to be easier but in the end that would still resort to Monte Carlo sums. Since model (2) is parametric, we could also compute distances between the parameters of the mixtures that depend on \mathbf{y} . That is for $k \in [K]$, between the mixing proportions $\eta_k^*(\mathbf{y}) = \frac{\pi_k^* \mathcal{N}(\mathbf{y}; \mathbf{c}_k^*, \boldsymbol{\Gamma}_k^*)}{\sum_{j=1}^K \pi_j^* \mathcal{N}(\mathbf{y}; \mathbf{c}_j^*, \boldsymbol{\Gamma}_j^*)}$ and conditional means $\mathbf{A}_k^* \mathbf{y} + \mathbf{b}_k^*$. 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 easily handled within ABC procedures.

Recently, developments regarding the Wasserstein distance have emerged ([Delon and Desolneux, 2020](#); [Chen et al., 2019](#)), introducing an optimal transport-based distance between Gaussian mixtures, denoted by MW_2 . The L_2 distance between mixtures is also straightforward to compute. Both distances are recalled in supplementary Section S2. We then derive two procedures respectively referred to as GLLiM-MW2-ABC and GLLiM-L2-ABC, writing sometimes GLLiM-D-ABC to include both cases and for generic distances D .

The semi-automatic ABC extensions that we propose are summarized in Algorithm 1. Algorithm 1 is presented with two simulated data sets, one for training GLLiM and constructing the surrogate posteriors, and one for the ABC procedure itself, but the same data set could be used. For rejection ABC, the selection also requires to fix a threshold ϵ . It is common practice to set ϵ to a quantile of the computed distances. GLLiM then requires the setting of K , the number of Gaussians in the mixtures, which can be chosen using model selection criteria (see [Deleforge et al., 2015b](#)). Its precise value is not critical, all the more so if GLLiM is not used for prediction, directly. See details in Section 6.

Algorithm 1 GLLiM-ABC algorithms – Vector and functional variants

-
- 415 **1: Inverse operator learning.** Apply GLLiM on a training set $\mathcal{D}_N =$
416 $\{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ to estimate, for any $\mathbf{z} \in \mathcal{Y}$, the K -Gaussian mixture
417 $p_G(\boldsymbol{\theta} | \mathbf{z}; \boldsymbol{\phi}_{K,N}^*)$ in (2) as a first approximation of the true posterior $\pi(\boldsymbol{\theta} | \mathbf{z})$,
418 where $\boldsymbol{\phi}_{K,N}^*$ does not depend on \mathbf{z} .
419
- 420 **2: Distances computation.** Consider another set $\mathcal{E}_M = \{(\boldsymbol{\theta}_m, \mathbf{z}_m), m \in [M]\}$.
421 For a given observed \mathbf{y} , do one of the following for $m \in [M]$:
422
- 423 **Vector summary statistics.** (Section 4.1)
424 GLLiM-E-ABC: Compute statistics $s_1(\mathbf{z}_m) = \mathbb{E}_G[\boldsymbol{\theta} | \mathbf{z}_m; \boldsymbol{\phi}_{K,N}^*]$ (3).
425 GLLiM-EV-ABC: Compute both $s_1(\mathbf{z}_m)$ and $s_2(\mathbf{z}_m)$ by considering also
426 posterior log-variances, *i.e.* the logarithms of the diagonal elements of (4).
427 In both cases, compute standard distances between summary statistics.
- 428 **Functional summary statistics.** (Section 4.2)
429 GLLiM-MW2-ABC: Compute $\text{MW}_2(p_G(\cdot | \mathbf{z}_m; \boldsymbol{\phi}_{K,N}^*), p_G(\cdot | \mathbf{y}; \boldsymbol{\phi}_{K,N}^*))$.
430 GLLiM-L2-ABC: Compute $\text{L}_2(p_G(\cdot | \mathbf{z}_m; \boldsymbol{\phi}_{K,N}^*), p_G(\cdot | \mathbf{y}; \boldsymbol{\phi}_{K,N}^*))$.
- 431 **3: Sample selection.** Select $\boldsymbol{\theta}_m$ values that lead to distances under an ϵ threshold
432 (rejection ABC) or apply an ABC procedure that can handle distances, directly.
- 433 **4: Sample use.** For a given observed \mathbf{y} , use the produced sample of $\boldsymbol{\theta}$ values to
434 compute a closer approximation of $\pi(\boldsymbol{\theta} | \mathbf{y})$.
-

5 Theoretical properties

436 Before illustrating the performance of GLLiM-D-ABC, we investigate the the-
437 oretical properties of our ABC quasi-posterior defined via surrogate posteriors.

438 Let $\mathcal{X} = \Theta \times \mathcal{Y}$ and $(\mathcal{X}, \mathcal{F})$ be a measurable space. Let λ be a σ -
439 finite measure on \mathcal{F} . Whenever we mention below that a probability measure
440 Pr on \mathcal{F} has a density, we will understand that it has a Radon–Nikodym
441 derivative with respect to λ (λ can typically be chosen as the Lebesgue
442 measure on a Euclidean space). For all $p \in [1, \infty)$ and f, g in appropriate
443 spaces, let $D_p(f, g) = (\int |f(\mathbf{x}) - g(\mathbf{x})|^p d\lambda(\mathbf{x}))^{1/p}$ denote the L_p distance and
444 $D_H^2(f, g) = \int (\sqrt{f(\mathbf{x})} - \sqrt{g(\mathbf{x})})^2 d\lambda(\mathbf{x})$ be the squared Hellinger distance. When
445 not specified otherwise, let D be an arbitrary distance on \mathcal{Y} or on densities,
446 depending on the context. We further denote the L_p norm for vectors by $\|\cdot\|_p$.

447 In a GLLiM-D-ABC procedure, the ABC quasi-posterior is constructed
448 as follows: let $p_G^{K,N}(\boldsymbol{\theta} | \mathbf{y}) = p_G(\boldsymbol{\theta} | \mathbf{y}; \boldsymbol{\phi}_{K,N}^*)$ be the surrogate conditional
449 distribution of form (2), learned from a preliminary GLLiM model with K com-
450 ponents and using a learning set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$. This conditional
451 distribution is a K -component mixture, which depends on a set of learned
452 parameters $\boldsymbol{\phi}_{K,N}^*$, independent of \mathbf{y} . The GLLiM-D-ABC quasi-posterior re-
453 sulting from the GLLiM-D-ABC procedure then depends both on K, N and
454 the tolerance level ϵ and can be written as

$$455 q_{G,\epsilon}^{K,N}(\boldsymbol{\theta} | \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(p_G^{K,N}(\cdot | \mathbf{y}), p_G^{K,N}(\cdot | \mathbf{z})) \leq \epsilon\}} f_{\boldsymbol{\theta}}(\mathbf{z}) d\mathbf{z}, \quad (5)$$

460

where D is a distance on densities such as the MW_2 and L_2 metrics, which are both proper distances (see supplementary Section S2).

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 to illustrate 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:

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y}, \mathbf{z}) \leq \epsilon\}} f_{\boldsymbol{\theta}}(\mathbf{z}) \, d\mathbf{z}. \quad (6)$$

If D is a distance and $D(\mathbf{y}, \mathbf{z})$ is continuous in \mathbf{z} , the ABC posterior in (6) 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 $\{\mathbf{z} \in \mathcal{Y} : D(\mathbf{y}, \mathbf{z}) \leq \epsilon\}$ in (6) tends to the singleton $\{\mathbf{y}\}$ so that consequently \mathbf{z} in the likelihood can be replaced by the observed \mathbf{y} , which leads to an ABC quasi-posterior proportional to $\pi(\boldsymbol{\theta})f_{\boldsymbol{\theta}}(\mathbf{y})$ and therefore equal 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 uses the equality, at convergence, of \mathbf{z} to \mathbf{y} , which is actually a stronger assumption than necessarily required for the result to hold. Alternatively, if we first rewrite (6) using Bayes' theorem, it follows that

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y}, \mathbf{z}) \leq \epsilon\}} \pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\mathbf{z}) \, d\mathbf{z} \propto \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y}, \mathbf{z}) \leq \epsilon\}} \pi(\boldsymbol{\theta} \mid \mathbf{z}) \pi(\mathbf{z}) \, d\mathbf{z}. \quad (7)$$

That is, when accounting for the normalizing constant:

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

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

507 following general result. Let us define our ABC quasi-posterior as,

508

509

510

511

512 which can be written as

513

514

515

516

517 The following theorem shows that $q_\epsilon(\cdot | \mathbf{y})$ converges to $\pi(\cdot | \mathbf{y})$ in total

518 variation, for fixed \mathbf{y} . The proof is detailed in supplementary Section S3.1.

519

520 **Theorem 1.** *For every $\epsilon > 0$, let $A_\epsilon = \{\mathbf{z} \in \mathcal{Y} : D(\pi(\cdot | \mathbf{y}), \pi(\cdot | \mathbf{z})) \leq \epsilon\}$.*

521 *Assume the following:*

522

523

524

525

526

527 *Under (A1)–(A4), $q_\epsilon(\cdot | \mathbf{y})$ in (9) converges in total variation to $\pi(\cdot | \mathbf{y})$, for*

528 *fixed \mathbf{y} , as $\epsilon \rightarrow 0$.*

529

530 It appears that what is important is not to select \mathbf{z} 's that are close (and

531 at the limit equal) to the observed \mathbf{y} but to choose \mathbf{z} 's so that the posterior

532 $\pi(\cdot | \mathbf{z})$ (the term appearing in the integral in (7)) is close (and at the limit

533 equal) to $\pi(\cdot | \mathbf{y})$. And this last property is less demanding than $\mathbf{z} = \mathbf{y}$.

534 Potentially, there may be several \mathbf{z} 's satisfying $\pi(\cdot | \mathbf{z}) = \pi(\cdot | \mathbf{y})$, but this

535 is not problematic when using (7), while it is problematic when following the

536 standard proof as in [Bernton et al. \(2019\)](#).

537

538 5.2 Convergence of the ABC quasi-posterior with

539

540 surrogate posteriors

541

542 In most ABC settings, based on data discrepancy or summary statistics, the

543 above consideration and result are not useful because the true posterior is

544 practically unknown and cannot be used to compare samples. However this

545 principle becomes useful in our setting, which is based on surrogate posteriors.

546 While the previous result can be seen as an oracle of sorts, it is more interesting

547 in practice to investigate whether a similar result holds when using surrogate

548 posteriors in the ABC likelihood. This is the goal of Theorem 2 below, which

549 we prove for a restricted class of target distribution and of surrogate posteriors

549 that are learned as mixtures.

550

551

552

552 We now assume that $\mathcal{X} = \Theta \times \mathcal{Y}$ is a compact set and consider the following

553 class $\mathcal{H}_{\mathcal{X}}$ of distributions on \mathcal{X} , $\mathcal{H}_{\mathcal{X}} = \{g_\varphi : \varphi \in \Psi\}$, with constraints on

554 the parameters, Ψ being a bounded parameter set. In addition the densities

in $\mathcal{H}_{\mathcal{X}}$ are assumed to satisfy the condition that for any $\varphi, \varphi' \in \Psi$ there exist arbitrary positive scalars a, b and B such that

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

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

$$\forall \boldsymbol{\theta} \in \Theta, \quad p^{K,N}(\boldsymbol{\theta} | \mathbf{z}) = p^K(\boldsymbol{\theta} | \mathbf{z}; \boldsymbol{\phi}_{K,N}^*),$$

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 $\pi(\cdot, \cdot)$:

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

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

$$q_{\epsilon}^{K,N}(\boldsymbol{\theta} | \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{A_{\epsilon, \mathbf{y}}^{K,N}}(\mathbf{z}) f_{\boldsymbol{\theta}}(\mathbf{z}) d\mathbf{z}. \quad (10)$$

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 $g_{\varphi} \in \mathcal{H}_{\mathcal{X}}$, $\pi(\mathbf{x}) = \int_{\Psi} g_{\varphi}(\mathbf{x}) G_{\pi}(d\varphi)$;
- (B2) The true posterior density $\pi(\cdot | \cdot)$ is continuous with respect to $\boldsymbol{\theta}$ and \mathbf{y} ;
- (B3) $D(\cdot, \cdot) : \Pi \times \Pi \rightarrow \mathbb{R}_+ \cup \{0\}$ is a metric on a functional class Π , which contains the class $\{p^{K,N}(\cdot | \mathbf{y}) : \mathbf{y} \in \mathcal{Y}, K \in \mathbb{N}^*, N \in \mathbb{N}^*\}$. In particular, $D(p^{K,N}(\cdot | \mathbf{y}), p^{K,N}(\cdot | \mathbf{z})) = 0$, if and only if $p^{K,N}(\cdot | \mathbf{y}) = p^{K,N}(\cdot | \mathbf{z})$;
- (B4) For every $\mathbf{y} \in \mathcal{Y}$, $\mathbf{z} \mapsto D(p^{K,N}(\cdot | \mathbf{y}), p^{K,N}(\cdot | \mathbf{z}))$ is a continuous function on \mathcal{Y} .

Then, under (B1)–(B4), the Hellinger distance $D_{\text{H}}(q_{\epsilon}^{K,N}(\cdot | \mathbf{y}), \pi(\cdot | \mathbf{y}))$ 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

$$\lim_{\epsilon \rightarrow 0, K \rightarrow \infty, N \rightarrow \infty} \Pr(\lambda(\{\mathbf{y} \in \mathcal{Y} : D_{\text{H}}^2(q_{\epsilon}^{K,N}(\cdot | \mathbf{y}), \pi(\cdot | \mathbf{y})) \geq \beta\}) \leq \alpha) = 1. \quad (11)$$

Sketch of the proof of Theorem 2.

For all $\boldsymbol{\theta} \in \Theta, \mathbf{y} \in \mathcal{Y}$, the quasi-posterior (10) can be written equivalently as

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

599
600
601

$$\text{with } K_\epsilon^{K,N}(\mathbf{z}; \mathbf{y}) = \frac{\mathbf{1}_{A_{\epsilon,\mathbf{y}}^{K,N}}(\mathbf{z}) \pi(\mathbf{z})}{\int_{\mathcal{Y}} \mathbf{1}_{A_{\epsilon,\mathbf{y}}^{K,N}}(\tilde{\mathbf{z}}) \pi(\tilde{\mathbf{z}}) d\tilde{\mathbf{z}}},$$

602
603
604
605
606

where $K_\epsilon^{K,N}(\cdot; \mathbf{y})$ 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 the Hellinger and L_1 distances (see details in supplementary Section S3.2 relations (28) and (29)), it then holds that

607
608

$$D_H^2(q_\epsilon^{K,N}(\cdot | \mathbf{y}), \pi(\cdot | \mathbf{y})) \leq 2D_H(\pi(\cdot | \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot | \mathbf{y})), \quad (12)$$

609
610

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

611
612
613

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

614

The next step is to bound the right-hand side of (12) 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 $\mathbf{z}_{0,\mathbf{y}}^{K,N} = \lim_{\epsilon \rightarrow 0} \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}$. 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 , $A_{0,\mathbf{y}}^{K,N} = \{\mathbf{z} \in \mathcal{Y} : D(p^{K,N}(\cdot | \mathbf{z}), p^{K,N}(\cdot | \mathbf{y})) = 0\}$ and $A_{0,\mathbf{y}}^{K,N} = \{\mathbf{z} \in \mathcal{Y} : p^{K,N}(\cdot | \mathbf{z}) = p^{K,N}(\cdot | \mathbf{y})\}$, using (B3). The distance on the right-hand side of (12) can then be decomposed in three parts,

622

623
624

$$\begin{aligned} D_H(\pi(\cdot | \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot | \mathbf{y})) &\leq D_H(\pi(\cdot | \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot | \mathbf{z}_{0,\mathbf{y}}^{K,N})) \\ &\quad + D_H(\pi(\cdot | \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}(\cdot | \mathbf{y})) \\ &\quad + D_H(p^{K,N}(\cdot | \mathbf{y}), \pi(\cdot | \mathbf{y})). \end{aligned} \quad (13)$$

625
626

627

628

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, and the definition of $\mathbf{z}_{0,\mathbf{y}}^{K,N}$ yields $p^{K,N}(\cdot | \mathbf{y}) = p^{K,N}(\cdot | \mathbf{z}_{0,\mathbf{y}}^{K,N})$.

632

Using that $\pi(\cdot | \cdot)$ is a uniformly continuous function in $(\boldsymbol{\theta}, \mathbf{y})$ on a compact set \mathcal{X} and taking the limit $\epsilon \rightarrow 0$, yields $\lim_{\epsilon \rightarrow 0} D_H^2(\pi(\cdot | \mathbf{z}_{\epsilon,\mathbf{y}}^{K,N}), \pi(\cdot | \mathbf{z}_{0,\mathbf{y}}^{K,N})) = 0$ 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 $\epsilon(\alpha_1, \beta_1) > 0$ such that for any $0 < \epsilon < \epsilon(\alpha_1, \beta_1)$,

637

638

639

640

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

641

642

643

644

Next, we prove that $D_H^2(\pi(\cdot | \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}(\cdot | \mathbf{y}))$ (which is equal to $D_H^2(\pi(\cdot | \mathbf{z}_{0,\mathbf{y}}^{K,N}), p^{K,N}(\cdot | \mathbf{z}_{0,\mathbf{y}}^{K,N}))$) and $D_H^2(p^{K,N}(\cdot | \mathbf{y}), \pi(\cdot | \mathbf{y}))$ both converge to 0 in measure λ , with respect to \mathbf{y} and in probability, with respect to \mathcal{D}_N .

Such convergence can be obtained via [Rakhlín et al. \(2005, Corollary 2.2\)](#), and Lemma 2 in supplementary Section S3.3.2, which provides the guarantee that we can choose a measurable function $\mathbf{y} \mapsto \mathbf{z}_{0,\mathbf{y}}^{K,N}$. Equation (11) in Theorem 2 follows from the triangle inequality (13). A detailed proof is provided in supplementary Section S3.2.

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 [Rakhlín 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

Let us recall that d is the observation dimension, ℓ the number of parameters and R the number of *i.i.d.* d -dimensional observations that may be available for each parameter value. We recall the notation $[N] = \{1, \dots, N\}$. Our first three examples are commonly used in the ABC literature and are there to illustrate the flexibility of our method, with an *i.i.d.* observation setting in Section 6.1 ($R = 100$, $d = 2$, $\ell = 2$) and Section 6.2 ($R = 100$, $d = 2$, $\ell = 5$), and a time series model ($R = 1$, $d = 150$, $\ell = 2$) in Section 6.3. For these examples, we compare with Wasserstein-ABC (WABC) of [Bernton et al. \(2019\)](#) using the **winference** R package ([Jacob et al., 2020](#)). WABC uses a SMC-ABC procedure instead of rejection ABC. When using SMC, we thus adopt the setting recommended in [Bernton et al. \(2019\)](#). In particular, the number of particles is set to 2048. In contrast, the other examples aim at departing from the usual benchmark examples in ABC. That is, we choose to consider settings that exhibit posterior distributions with characteristics such as multimodality and heavy tails. We report a synthetic experiment where the posterior distribution has mass on four 1D manifolds (Section 6.4). Other synthetic examples are described in supplementary Section S4.4. All these other 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 6.5.

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, was the posterior mean. This posterior mean is not

691 known and needs to be approximated, *e.g.* by linear regression. In this section,
692 the transformations used for the regression part are $(1, y, y^2, y^3, y^4)$ following
693 the procedure suggested in the **abctools** package (Nunes and Prangle, 2015).
694 We refer to this procedure as semi-automatic ABC. This approach using the
695 posterior mean approach is further developed in Jiang et al. (2017), where a
696 multilayer perceptron deep neural network regression model is employed. The
697 deep neuronal network with multiple hidden layers considered by Jiang et al.
698 (2017) offers stronger representational power to approximate the posterior
699 mean and hence to learn an informative summary statistic, when compared to
700 linear regression models. Improved results were obtained by Jiang et al. (2017),
701 but we did not compare our approach to their method, except by reporting
702 some of their results when relevant. Discrepancy-based results from Nguyen
703 et al. (2020a) are also reported when available.

704 The performances of the four proposed GLLiM-ABC schemes summarized
705 in Algorithm 1 are compared to that of semi-automatic ABC. When not spec-
706 ified otherwise, reported results are obtained with a simple rejection scheme
707 as per instances implemented in the **abc** R package (Csillery et al., 2012).
708 The other schemes available in the **abc** package have been tested but no no-
709 table performance differences were observed. In regards to the final sample
710 thresholding (*i.e.*, choice of ϵ), following common practice, all methods retain
711 samples for which the distance to the observation is under a small (*e.g.* 0.1%)
712 quantile of all computed distances. Alternatively, we also report results with
713 a SMC-ABC scheme as implemented in the **winference** package.

714 The **xLLiM** R package (Perthame et al., 2017), available on the CRAN, is
715 used to learn a GLLiM model with K components from a set \mathcal{D}_N of N simula-
716 tions from the true model, meaning that each pair $(\boldsymbol{\theta}_n, \mathbf{y}_n)$ in \mathcal{D}_N is obtained
717 by simulating $\boldsymbol{\theta}_n$ from the prior $\pi(\boldsymbol{\theta})$ and \mathbf{y}_n from the likelihood $f_{\boldsymbol{\theta}_n}(\mathbf{y})$. The
718 selection of K using the Bayesian Information Criterion (BIC) is illustrated in
719 Sections 6.3 to 6.5. The GLLiM implementation uses an isotropic constraint
720 except for the first three examples as specified below. The isotropic GLLiM in-
721 volves less parameters than the fully-specified GLLiM and we observed that, in
722 the one observation settings, it yielded surrogate posteriors of sufficient qual-
723 ity for the ABC selection scheme. The exact meaning of this constraint can be
724 found in Deleforge et al. (2015b). Another set of simulated pairs $(\boldsymbol{\theta}, \mathbf{y})$ of size
725 M is generally used for the ABC scheme unless otherwise specified.

726 To visualize posterior samples densities, we use a density estimation
727 procedure based on the **ggplot2** R package with a Gaussian kernel.

728 Computing times for the various procedures and experiments are discussed
729 in Section 6.6 and shown in Table S3 in supplementary Section S5.

730

731 6.1 Normal Location model

732

733 Our first illustrations correspond to situations where, for each possible value
734 of the parameter, it is possible to simulate or observe many (R) *i.i.d.* realiza-
735 tions. The observations to be inverted are also made of R *i.i.d.* realizations but
736 assuming a different number is not a problem.

We first consider the normal location model described in Section 2.2 of [Bernton et al. \(2019\)](#). This model is a particular case of the following model. In the bivariate case, the parameter is a 2-dimensional vector $\boldsymbol{\theta}$, which is assigned a Gaussian prior $\mathcal{N}_2(\cdot; \mathbf{c}, \boldsymbol{\Gamma})$ with mean \mathbf{c} and covariance matrix $\boldsymbol{\Gamma}$. The observed variable \mathbf{y} is then assumed to follow a Gaussian distribution $\mathcal{N}_2(\cdot; \mathbf{A}\boldsymbol{\theta} + \mathbf{b}, \boldsymbol{\Sigma})$. The example of [Bernton et al. \(2019\)](#) corresponds to $\mathbf{c} = 0$, $\boldsymbol{\Gamma} = 25\mathbf{I}$, $\mathbf{A} = \mathbf{I}$, $\mathbf{b} = 0$ and $\boldsymbol{\Sigma}$ is equal to 1 on the diagonal and 0.5 off the diagonal. For comparison with their WABC procedure, we use the exact setting described in this paper. A sample $\{\mathbf{y}^r, r \in [R]\}$ of $R = 100$ *i.i.d.* observations is generated from a bivariate normal distribution. The mean components are drawn from a standard normal distribution, and the values generated are approximately -0.71 and 0.09 . For this model, the posterior is available in closed form and is Gaussian. Details can be found in supplementary Section S4.1. This normal location model is exactly the GLLiM model for $K = 1$ and is therefore a particularly favorable example for our procedures. Although the example may be simplistic, the availability of the true posterior distribution and closed-form expressions for the distances provides some interesting insights into our proposed approach and how it differs and compares to the WABC approach of [Bernton et al. \(2019\)](#). We report in supplementary Section S4.1 results for an SMC-ABC algorithm using GLLiM successively with the MW_2 and L_2 distance and the Wasserstein distance between samples (WABC). Despite its simplicity, this example clearly shows the difference between the L_2 and the Wasserstein distances. In this example, the MW_2 and L_2 distances are explicit functions of the difference between the sufficient sample means while the Wasserstein distance of WABC measures the difference between sample histograms. However, we suspect the exponential form in the L_2 distance generates a very specific behaviour compared to the other distances (see supplementary Section S4.1 for details).

Overall, the GLLiM-based procedures are more efficient in terms of simulations and time (See supplementary Figure S1 and Table S3) but note that this can be very specific to this example, which simplifies the expressions of our distances greatly, while the cost of computing a Wasserstein distance between samples (WABC) does not depend on the model under consideration but only on the observations dimension and number. Also it appears that the L_2 distance requires more simulations to be as efficient as MW_2 .

6.2 Bivariate Beta model

In contrast to the previous example, the bivariate Beta model is a typical target for ABC procedures as nor the likelihood neither the posterior distribution are available in closed-form or obtained via another reference procedure. This is problematic to assess the quality of the posterior approximations. We thus follow the analysis done in most ABC papers (*e.g.* [Crackel and Flegal \(2017\)](#); [Bernton et al. \(2019\)](#); [Nguyen et al. \(2020a\)](#); [Jiang et al. \(2018\)](#), etc.), which mainly report the concentration of the posterior approximations around the data-generating parameters. Note that a number of potential metrics have been

783 listed in [Lueckmann et al. \(2021\)](#) but they are not practical for comparing
784 samples produced by ABC schemes and are computationally costly.

785 The bivariate Beta model proposed by [Crackel and Flegal \(2017\)](#) and also
786 used by [Nguyen et al. \(2020a\)](#); [Jiang et al. \(2018\)](#) is defined with five positive
787 parameters $\theta_1, \dots, \theta_5$ by letting $v_1 = (u_1 + u_3)/(u_5 + u_4)$ and $v_2 = (u_2 +$
788 $u_4)/(u_5 + u_3)$, where $u_i \sim \text{Gamma}(\theta_i, 1)$, for $i \in [5]$, and setting $z_1 = v_1/(1 +$
789 $v_1)$ and $z_2 = v_2/(1 + v_2)$. The likelihood for the bivariate random variable
790 $\mathbf{z}^\top = (z_1, z_2)$ is not available in closed form. The observed sample is generated
791 from the model with values $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = (1, 1, 1, 1, 1)$. The prior on each
792 parameter is taken to be independent and uniform over interval $[0, 5]$.

793 We fit a GLLiM model with $K = 100$ for *i.i.d.* data (see Section S2.1 in
794 supplementary material) to a set made of $N = 10^5$ 5-dimensional vectors of
795 parameters, each associated to $R = 100$ *i.i.d.* bivariate observations.
796

797 6.2.1 Comparison of rejection ABC procedures

798 We first use this same set for a rejection ABC approach with a tolerance
799 threshold ϵ set to the 0.05% quantile leading to selected samples of size 50, in
800 order to match the experiments of [Nguyen et al. \(2020a\)](#); [Jiang et al. \(2018\)](#).

801 The marginal ABC posterior distributions of parameters $\theta_1, \theta_2, \theta_3, \theta_4$ and
802 θ_5 are displayed in Figure S2 of the supplementary material. Results are qual-
803 itatively similar to that of [Nguyen et al. \(2020a\)](#); [Jiang et al. \(2018\)](#), which
804 use data discrepancies. Our GLLiM-ABC procedures can be seen as direct al-
805 ternatives to these latter methods. In contrast, to apply semi-automatic ABC
806 requires summary statistics. In absence of candidate summary statistics, it is
807 suggested by [Fearhead and Prangle \(2012\)](#) to use evenly-spaced quantiles. For
808 comparison, following [Jiang et al. \(2018\)](#), we apply the semi-automatic proce-
809 dure on 7 quantiles from the first observed dimension and 7 quantiles from the
810 second. Each simulated data set of size $2 \times R$ is then reduced to 14 quantiles.
811

812 Although the use of somewhat arbitrary summary statistics is often prob-
813 lematic, we observe that using 14 quantiles in this case provides reasonable
814 results. Visually (see Figures S2 and S3 in the supplementary material), semi-
815 automatic ABC shows modes close to the data-generating parameter values.
816 The GLLiM mixture appears to provide slightly shifted modes that are closer
817 located after an ABC step is added, except for GLLiM-L2-ABC. In this exam-
818 ple, the L_2 distance shows quite different posterior shapes. Overall the results
819 are qualitatively similar to that in [Jiang et al. \(2018\)](#).

820 For a more complete comparison, we also apply the other GLLiM-ABC
821 methods with the 14 quantiles summaries. The standard GLLiM implemen-
822 tation is used with $K = 40$ and no constraint. Our GLLiM-ABC procedures
823 easily apply in this new setting, while the discrepancy-based methods de-
824 scribed in [Bernton et al. \(2019\)](#); [Jiang et al. \(2017\)](#); [Nguyen et al. \(2020a\)](#)
825 are not designed for this situation. Supplementary Figure S3 shows marginal
826 posteriors for the 5 parameters and 5 procedures. GLLiM-MW2-ABC and
827 GLLiM-E-ABC perform similarly, while the addition of log-variances in
828 GLLiM-EV-ABC does not seem to effect the posterior shapes, significantly. In

contrast, GLLiM-L2-ABC performs very differently with modes further away from the data-generating values.

For a more quantitative comparison, we compute for each posterior samples of size S , empirical means of the parameters, $\bar{\theta}_j = \frac{1}{S} \sum_{i=1}^S \theta_j^i$, and empirical root mean square errors (RMSE) defined as $R(\theta_j) = \sqrt{\frac{1}{S} \sum_{i=1}^S (\theta_j^i - \theta_j^0)^2}$ where $j \in [5]$, $S = 50$ and θ_j^i is the sample i for θ_j and θ_j^0 is the true parameter value. Table 1 shows these quantities averaged over 10 repetitions of the same experiment. The RMSE reported in Table 1 confirm that semi-automatic ABC when using quantiles as summary statistics and GLLiM-MW2-ABC method in both cases, with or without summary statistics, provide posterior approximations more concentrated around the data-generating parameter values. Overall, all methods have similar performance except for GLLiM-L2-ABC. Since our setting is the same as in Nguyen et al. (2020a), we also show in Table 1 the best results obtained for this example, adapted with only $R = 100$ *i.i.d.* observations instead of $R = 500$ originally in Nguyen et al. (2020a). Although a different set of simulations has been used and the results are not strictly comparable, our results are qualitatively similar to that of Nguyen et al. (2020a).

6.2.2 SMC-ABC and comparison with WABC

We then consider SMC-ABC as an alternative to rejection ABC. To compare with the WABC approach of Bernton et al. (2019), we use the SMC-ABC implementation proposed in this paper. This SMC setting being quite different, in terms of tuning requirements, the comparison is made on another set of simulations, with a similar budget. Specifically, we consider a first budget of $M = 10^5$ as before and a larger one of $M = 10^6$. The SMC-ABC is run with these respective budgets following the recommendations of Bernton et al. (2019). The number of particles is set to 2048, which is also the size of the retained ABC samples. The resulting posterior approximations are shown in supplementary Figure S4.

As already mention, we cannot make conclusions regarding the proximity to the true posterior distribution. However, it appears clearly that a higher budget tends to concentrate the posterior approximations closer to the data-generating values, and this more significantly so for GLLiM-MW2-SMC-ABC and WABC while GLLiM-L2-SMC-ABC does not always concentrate at the same location. We have not further investigated the reasons for this latter different behaviour but it may be related to what we had already observed in the simpler normal location model case (see supplementary Figure S1). For the L_2 distance, SMC-ABC shows more numerical difficulties, *e.g.* with smaller acceptance ratios at each step (around 35%). Supplementary Table S1 summarizes the comparison.

Table 1 Bivariate Beta model: Empirical parameter means, and RMSE for ABC posterior samples averaged over 10 repetitions of the experiment with observed data generated with $\theta = (1, 1, 1, 1, 1)$. The ABC posterior values are computed as empirical values over samples of size 50. Average means closest to 1 and best (lowest) average RMSE values are in boldface. The best results obtained by the approach of [Nguyen et al. \(2020a\)](#) using various data discrepancies, in the same setting ($R = 100$) but with a different set of simulations, are also provided for comparison.

Procedure	$\bar{\theta}_1$	$\bar{\theta}_2$	$\bar{\theta}_3$	$\bar{\theta}_4$	$\bar{\theta}_5$	R(θ_1)	R(θ_2)	R(θ_3)	R(θ_4)	R(θ_5)
GLLiM mixture	2.510	2.546	2.714	2.630	2.591	2.145	2.291	2.201	2.277	2.056
GLLiM-E-ABC	1.439	1.051	0.914	1.095	1.264	0.952	0.791	0.483	0.629	0.510
GLLiM-EV-ABC	1.444	1.037	0.916	1.153	1.205	1.003	0.751	0.556	0.596	0.521
GLLiM-L2-ABC	1.860	2.301	2.430	2.136	2.620	1.268	1.859	2.008	1.536	1.966
GLLiM-MW2-ABC	1.330	1.000	0.8465	1.056	1.159	0.836	0.781	0.458	0.558	0.448
with 14 quantiles as summaries										
Semi-auto ABC	1.235	1.173	0.948	1.000	1.145	0.7601	0.747	0.597	0.599	0.582
GLLiM mixture	0.922	1.139	1.002	0.917	1.040	1.869	1.802	1.286	1.231	0.993
GLLiM-E-ABC	1.209	1.438	1.146	1.071	1.302	0.699	0.880	0.632	0.597	0.659
GLLiM-EV-ABC	1.215	1.565	1.157	1.084	1.167	0.748	0.999	0.677	0.660	0.599
GLLiM-L2-ABC	3.339	2.989	3.420	3.315	2.601	2.711	2.462	2.655	2.715	1.958
GLLiM-MW2-ABC	1.159	1.460	1.146	1.079	1.264	0.687	0.877	0.607	0.593	0.634
Best results using data discrepancies as in Nguyen et al. (2020a)										
$R = 100$	1.275	1.176	0.751	0.830	1.237	0.834	0.593	0.459	0.219	0.409

6.3 Moving average model

The moving average model is widely used in time series analysis. In particular the moving average model of order 2, MA(2), has often illustrated ABC procedures ([Marin et al., 2012](#); [Jiang et al., 2018, 2017](#); [Fearnhead and Prangle, 2012](#); [Nguyen et al., 2020a](#)). Natural summary statistics are the empirical auto-covariances of lag 1 and 2. This example is a way to illustrate our method on time series in the same manner as [Bernton et al. \(2019\)](#). In contrast to the previous example, we consider that we have a single observation which is a time series of length d . However, we treat it as a set of *i.i.d.* observations of smaller length. This corresponds to the approximation suggested in Section 4.2 of [Bernton et al. \(2019\)](#). Their Wasserstein-ABC proposal uses empirical distributions and, like other data discrepancy based methods, is in principle only valid for *i.i.d.* observations. However, they also investigate the use of the method to time series where observations are not *i.i.d.*. We make a similar attempt in this work and show how it can be interpreted in our framework. To favor comparison with other results on the MA(2) model, we adopt a similar setting as in most papers, *i.e.* that of [Jiang et al. \(2017\)](#), but a quantitative comparison is not strictly possible as the simulated observations may vary from one paper to another. The MA(2) process is a stochastic process $(y'_t)_{t \in \mathbb{N}^*}$ defined by

$$y'_t = z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2}, \quad (14)$$

where $\{z_t\}$ is an *i.i.d.* sequence, according to a standard normal distribution and θ_1 and θ_2 are scalar parameters. A standard identifiability condition is imposed on this model leading to a prior distribution on the triangle described by the inequalities $-2 < \theta_1 < 2$, $\theta_1 + \theta_2 > -1$, $\theta_1 - \theta_2 < 1$. The prior on the two model parameters is taken uniform over the triangular domain. For

each pair of parameters (θ_1, θ_2) in the triangular domain, a series of length 150 is simulated according to model (14). This is repeated $N = 10^5$ times. The series to be inverted is simulated similarly with true parameters $\theta_1 = 0.6$ and $\theta_2 = 0.2$. For ABC procedures, the tolerance threshold ϵ is set to the 0.1% quantile leading to selected samples of size 100.

To learn a GLLiM model with $d = 150$, $\ell = 2$, we propose to use the *i.i.d.* adaptation of GLLiM (see supplementary material S1.2). In terms of GLLiM, this is equivalent to assume block diagonal covariance matrices when approximating the likelihood. There is some flexibility as regards the block sizes. Larger blocks depart less from the true MA(2) model while requiring more parameters to be estimated. Smaller blocks correspond to neglect some of the dependencies between the blocks but may be acceptable if the remaining dependencies carry enough information on the parameters. Two block decompositions are tested. All series of length 150 (y_1, \dots, y_{150}) are first cut into $R = 50$ smaller series of length 3, $(y_1, y_2, y_3), (y_4, y_5, y_6), \dots$, which are considered as independent and identically distributed. GLLiM is applied with $d = 3$, $R = 50$ and no constraint on the 3×3 blocks themselves. A second experiment is made with $R = 5$ and $d = 30$ *i.e.* with 5 unconstrained blocks of size 30×30 . A better precision especially on θ_2 is obtained with this later setting. This confirms the sensitivity of the dependence over time information in the MA(2) model. We thus choose this setting considering each time series as a sample of 5 smaller series of length 30. To illustrate the possibility to select the number of GLLiM components K in a more data-driven way, we compute the Bayesian Information Criterion (BIC) for $K = 2$ to 30. The value of K leading to the minimum BIC is then selected. The supplementary Figure S5 shows the BIC values, which flattens after $K = 15$ and whose minimum is for $K = 20$. We therefore use a GLLiM model learned with $K = 20$. For comparison posterior samples obtained with $K = 30$ are also shown in supplementary Figure S6. The results are similar for both values of K without a clear difference in favor of the selected K . Figure S6 also shows samples obtained with WABC and GLLiM using SMC-ABC instead of Rejection ABC. WABC performs poorly (Figure S6 (m)) due to the low $R = 5$ (see also Table S2).

We also compare with semi-automatic ABC applied directly to the time series of length 150. Reducing the time series into smaller time series is not possible as the approach is not designed to handle *i.i.d.* observations. Instead we also consider the two empirical auto-covariances as summary statistics. Empirical values for parameter means, standard deviations and correlation, when applying the different ABC schemes for one observed time series, are compared to the true ones computed numerically with importance sampling. The corresponding ABC estimations and samples are shown in supplementary material Table S2 and Figure S6. The results are qualitatively similar to that of Jiang et al. (2017) with a poor estimation of the means for semi-automatic ABC on the full time series. They also confirm results already observed in previous works, namely that semi-automatic and auto-covariance-based procedures do not well capture correlation information between θ_1 and θ_2 .

967 We then repeat the comparison for 100 different observed series, all simu-
968 lated from true parameters (0.6, 0.2). In each case, the true posterior means,
969 standard deviations of θ_1 and θ_2 , and correlation are computed numerically.
970 The mean squared errors (MSE) to the true posterior values are then com-
971 puted and reported in Table 2. These values are computed using selected
972 samples of size 100 each. The first line in Table 2 shows the averages over
973 the 100 experiments of the posterior true quantities, numerically computed.
974 In particular, we see that the averaged posterior means get close to the true
975 values 0.6 and 0.2. Most results correspond to a rejection ABC procedure. For
976 comparison, we also give the MSE obtained with a SMC-ABC implementa-
977 tion for a GLLiM-MW2 distance (referred to as simply GLLiM-MW2-SMC
978 for a shorter name). As before SMC is run with 2048 particles but MSE are
979 computed by selecting the parameters values corresponding to the best 100
980 distances among the 2048. WABC is not further tested due to its poor perfor-
981 mance in this example. Two sets of results are given corresponding respectively
982 to $K = 20$ and $K = 30$. The $K = 30$ best results are slightly better. This
983 may be due to a better model fit, while selecting K using BIC also accounts
984 for model complexity. For $K = 20$, the best MSE are obtained with GLLiM-
985 MW2-SMC and GLLiM-MW2-ABC except for the correlation MSE which is
986 best for GLLiM-EV-ABC. Semi-automatic ABC applied directly on the time
987 series provides the largest errors. Semi-automatic ABC provides much lower
988 errors when applied on auto-covariances. The methods using auto-covariances
989 provide satisfying results for the θ_1 mean but not for the other quantities. The
990 GLLiM mixture provides better estimates than semi-automatic ABC on the
991 full time series but remains far from the best performance. This illustrates
992 again that there is a clear gain in complementing GLLiM with an ABC step
993 and that the initial GLLiM mixture needs not to be very accurate. The second
994 best method is GLLiM-L2-ABC, which performs similarly as GLLiM-E-ABC,
995 while surprisingly adding the log-variances in GLLiM-EV-ABC seems to de-
996 grade the performance except for the correlation. This illustrates the fact that
997 in this unimodal posterior case, the posterior expectation is a good summary
998 statistic. Note however, that GLLiM-MW2-ABC still provides a performance
999 gain. To compare with another method that uses estimates of posterior expec-
1000 tations as summary statistics, we report results given in [Jiang et al. \(2017\)](#).
1001 Their deep neural network-based method (DNN) provides larger MSE than
1002 our GLLiM-ABC methods.

1003

1004 6.4 Multiple hyperboloid example

1005

1006 Our main targets are posterior distributions with multiple modes for which our
1007 method is more likely to provide significantly better performance than existing
1008 approaches. It is straightforward to construct models that lead to multimodal
1009 posteriors by considering likelihoods that are invariant by some transformation.
1010 Such non-identifiable models include ill-posed inverse problems that can be
1011 constructed as explained in Section S4.4 of the supplementary material. Three

1012

Table 2 MA(2) model: mean squared errors (MSE) over 100 simulated observations with the same true parameters (0.6,0.2). MSE are computed for all methods, for the estimated parameter means, standard deviations and correlations compared to their true counterparts computed numerically. Three sets of results are shown, corresponding to procedures that does not used GLLiM, procedures using GLLiM learned with $K = 30$ and $K = 20$ components. The last line shows values as reported in [Jiang et al. \(2017\)](#) based on a deep neural network learning (DNN). The "Exact" line reports the means of the 100 true posterior values. Best (lowest) MSE values are in boldface with a * to indicate the overall best values.

Procedure	mean(θ_1)	mean(θ_2)	std(θ_1)	std(θ_2)	cor(θ_1, θ_2)
Average					
Exact	0.5807	0.1960	0.0810	0.0813	0.4483
MSE					
Semi-auto ABC	0.3402	0.0199	0.1521	0.1255	0.2235
Auto-cov Semi-auto	0.0048	0.0147	0.0012	0.0070	0.1212
Auto-cov Rejection ABC	0.0047	0.0145	0.0010	0.0070	0.1196
$K = 30$					
GLLiM mixture	0.0142	0.0046	0.1652	0.0399	0.1734
GLLiM-E-ABC	0.0040	0.0039	0.0005	0.0003	0.0446
GLLiM-EV-ABC	0.0060	0.0040	0.0035	0.0014	0.0632
GLLiM-L2-ABC	0.0037	0.0041	0.0005	0.0005	0.0501
GLLiM-MW2-ABC	0.0027*	0.0021*	0.0002*	0.0003*	0.0356*
$K = 20$					
GLLiM mixture	0.0340	0.0060	0.1223	0.0367	0.1691
GLLiM-E-ABC	0.0103	0.0066	0.0020	0.0037	0.0440
GLLiM-EV-ABC	0.0256	0.0065	0.0052	0.0035	0.0375
GLLiM-L2-ABC	0.0095	0.0057	0.0016	0.0031	0.0470
GLLiM-MW2-ABC	0.0038	0.0041	0.0005	0.0013	0.0509
GLLiM-MW2-SMC	0.0032	0.0035	0.0003	0.0010	0.0513
ABC-DNN Jiang et al. (2017)	0.0096	0.0089	0.0025	0.0026	0.0517

synthetic examples therein show that the expectation as a summary statistic suffers from the presence of two equivalent modes, while GLLiM-D-ABC procedures well capture multimodality.

In this sub-section, we consider a more complex non-identifiable example constructed from a real sound source localization problem in audio processing. This example is artificial. The link to audio processing is only illustrative and further detail is provided in supplementary Section S4.5.

The object of interest is an unknown parameter $\theta = (x, y)$ that can be interpreted as a source location in a 2D scene. To create a multimodal posterior, we consider the following likelihood that depends on two pairs $\mathbf{m}^1 = (\mathbf{m}_1^1, \mathbf{m}_2^1)$ and $\mathbf{m}^2 = (\mathbf{m}_1^2, \mathbf{m}_2^2)$ of 2-dimensional parameters. We assume a d dimensional observation $\mathbf{y} = (y_1, \dots, y_d)$ with

$$f_{\theta}(\mathbf{y}) = \frac{1}{2} \mathcal{S}_d(\mathbf{y}; F_{\mathbf{m}^1}(\theta) \mathbb{I}_d, \sigma^2 \mathbb{I}_d, \nu) + \frac{1}{2} \mathcal{S}_d(\mathbf{y}; F_{\mathbf{m}^2}(\theta) \mathbb{I}_d, \sigma^2 \mathbb{I}_d, \nu), \quad (15)$$

$$\text{where } F_{\mathbf{m}}(\theta) = (\|\theta - \mathbf{m}_1\|_2 - \|\theta - \mathbf{m}_2\|_2), \text{ if } \mathbf{m} = (\mathbf{m}_1, \mathbf{m}_2). \quad (16)$$

The above likelihood corresponds to a mixture with equal weight of two d -variate Student t -distributions with a d -dimensional location parameter with

1059 all dimensions equal to $F_{\mathbf{m}^1}(\boldsymbol{\theta})$ (resp. $F_{\mathbf{m}^2}(\boldsymbol{\theta})$), diagonal isotropic scale matrix
 1060 equal to $\sigma^2 \mathbf{I}_d$ and degree-of-freedom (dof) parameter ν .

1061 The parameter space is assumed to be $\Theta = [-2, 2] \times [-2, 2]$ and the
 1062 prior on $\boldsymbol{\theta}$ is assumed to be uniform on Θ . The pair positions are $\mathbf{m}^1 =$
 1063 $((-0.5, 0), (0.5, 0))$ and $\mathbf{m}^2 = ((0, -0.5), (0, 0.5))$. We assume $\nu = 3$ and
 1064 $\sigma^2 = 0.01$. The true $\boldsymbol{\theta}$ is set to $\boldsymbol{\theta} = (1.5, 1)$ and we simulate a 10-dimensional \mathbf{y}
 1065 following model (15). Depending on whether this observation is coming from
 1066 the first pair or second pair component, it results a true posterior as shown
 1067 in Figure 1 (d) or one with non-intersecting hyperbolas. The contour plot
 1068 indicates that the observation corresponds to the $((0, -0.5), (0, 0.5))$ pair. Mul-
 1069 timodality of the posterior is coming from that each isosurface defined by (16)
 1070 is represented by a two-sheet hyperboloid in 2D.

1071 The four ABC methods using GLLiM and semi-automatic ABC are com-
 1072 pared. The first GLLiM model used consists of $K = 20$ Gaussian components
 1073 with an isotropic constraint. A selected sample of 1000 values is retained
 1074 by thresholding the distances under the 0.1% quantile. In a first test, semi-
 1075 automatic ABC and GLLiM use the same data set of size $M = 10^6$, which
 1076 is also used for the rejection ABC part. Selected samples are shown in sup-
 1077 plementary Section S4.5.2, Figure S10. The mixture provided by GLLiM as
 1078 an approximation of the true posterior (Figure 10 (d)) well captures the main
 1079 posterior parts. This GLLiM posterior is a 20-component Gaussian mixture of
 1080 form (2). The true posterior expectations are all zero and are thus not informa-
 1081 tive about the location parameters. However, a correct structure can be seen in
 1082 the GLLiM-E-ABC sample, in contrast to the semi-automatic one that shows
 1083 no structure as expected. Adding the posterior log-variance estimations has a
 1084 good impact on the selected sample, which is only marginally different from
 1085 the GLLiM-D-ABC samples. This suggests that the posterior log-variances are
 1086 very informative on the location parameters.

1087 When GLLiM is first learned with a smaller data set of size $N = 10^5$ and
 1088 different from the rejection ABC data set, results slightly degrade, but not
 1089 significantly so (Supplementary Figure S11). More badly localized estimations
 1090 can be seen in the samples of Figure S11 (g,h), but the GLLiM-D-ABC samples
 1091 are well localized and are not really impacted by this difference in the GLLiM
 1092 learning step. In this case the improvement of GLLiM-D-ABC over GLLiM-
 1093 EV-ABC is clearer.

1094 When BIC is used to select K , we observe a minimum at $K = 38$ when
 1095 the criterion is computed for $K = 2$ to $K = 40$ (see supplementary Figure
 1096 S12). Figure 1 below shows then the results with GLLiM learned with $K =$
 1097 38 and $N = 10^5$. A clear improvement is visible especially on the GLLiM-
 1098 mixture and GLLiM-EV-ABC plots. In contrast to the MA(2) example where
 1099 manually choosing K too large led to similar results, choosing it too small
 1100 has here more impact. We also use the better GLLiM approximation to show
 1101 that the number of ABC simulations can be reduced without much changing
 1102 the selected posterior samples. Plots (c) and (g) in Figure 1 are obtained by
 1103 selecting among $M = 10^5$ simulations the best 1% distances instead of the best
 1104

0.1% in supplementary Figure S11. At last, all previously mentioned samples are obtained using a rejection ABC scheme while Figure 1 (h) is a sample obtained using the MW_2 distance and SMC-ABC. Results are very similar with a slightly better sampling with SMC at the hyperboloids intersection.

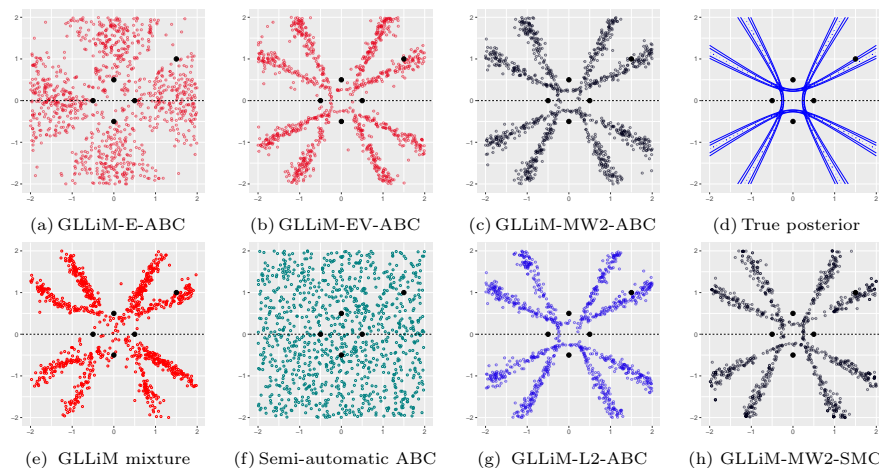


Figure 1 Multiple hyperboloid example. GLLiM is learned with $K = 38$ on a data set of size $N = 10^5$ while ABC is run using a data set of size $M = 10^6$ for (a,b,f,h) and $M = 10^5$ for (c,g). Rejection ABC is used except for (h) which uses SMC-ABC. Selected samples using (a) GLLiM posterior expectations, (b) GLLiM posterior expectations and log variances, (c) MW_2 distances, (d) contours of the true posterior distribution, (e) approximate GLLiM posterior for the observed data, (f) semi-automatic ABC, (g) L_2 distances and (h) MW_2 distances with SMC-ABC. Black points on the dotted line are the pairs positions. The fifth black point is the true parameter values.

6.5 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, photometric 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 planet Mars is described by [Murchie et al. \(2009\)](#); [Fernando et al. \(2016\)](#). Such observations can also be measured in the laboratory, on known materials 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 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

1151 $\boldsymbol{\theta} \in \mathbb{R}^4$ to a *theoretical* BRDF denoted by $\mathbf{y} = F_{\text{Hapke}}(\boldsymbol{\theta}) \in \mathbb{R}^d$. A given exper-
 1152 iment defines d geometries of measurement, each parameterized by a triplet
 1153 (θ_0, θ, ϕ) of incidence, emergence and azimuth angles. Moreover, $\boldsymbol{\theta} = (\omega, \bar{\theta}, b, c)$
 1154 are the sensitive parameters, respectively single scattering albedo, macroscopic
 1155 roughness, asymmetry parameter and backscattering fraction. More details on
 1156 these quantities and their photometric meanings may be found in [Schmidt and](#)
 1157 [Fernando \(2015\)](#); [Labarre \(2017\)](#). Although available, the expression of F_{Hapke}
 1158 is very complex and tedious to handle analytically, with a number of approxi-
 1159 mations required (see the description of the function in more than 15 pages in
 1160 [Labarre 2017](#)). In practice, it is therefore mainly used via a numerical code, al-
 1161 lowing simulations from the model. In addition, previous studies ([Kugler et al.](#)
 1162 [2021](#); [Schmidt and Fernando 2015](#)) have shown evidence for the existence of
 1163 multiple solutions or for the possibility to obtain very similar observations from
 1164 different sets of parameters, which makes this setting appropriate for testing
 1165 the ability of our procedures to recover multimodal posterior distributions.

1166 In the following experiments, all parameters are transformed to be in $[0, 1]^4$,
 1167 which amounts to keep b and c unchanged, divide $\bar{\theta}$ by 30 and operate the
 1168 following change of variable for ω , $\gamma = 1 - \sqrt{1 - \omega}$. This last transformation
 1169 also has the advantage of avoiding the non-linearity of F_{Hapke} , when ω tends to
 1170 1. The experimental setting defines geometries at which the measurements are
 1171 made, which in turn define F_{Hapke} . The number of geometries thus corresponds
 1172 to the size d , of each observation. The measurement geometries used to define
 1173 F_{Hapke} are borrowed from a real laboratory experiment presented below. The
 1174 number of parameters is therefore $\ell = 4$ with $d = 10$ observed geometries.
 1175 The sets to learn GLLiM and generate ABC samples are both set to size
 1176 $N = M = 10^5$. For each pair $(\boldsymbol{\theta}, \mathbf{y})$ in the simulated data sets, the 4 parameters
 1177 $(\boldsymbol{\theta})$ are simulated uniformly in $[0, 1]^4$. Besides these learning sets, the Hapke
 1178 simulator is not available to us so that we cannot run SMC-ABC for this specific
 1179 example. Following a previous study ([Kugler et al., 2021](#)), the corresponding
 1180 reflectance curves are generated as $\mathbf{y} = F_{\text{Hapke}}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a centered
 1181 Gaussian variable with isotropic covariance $\sigma^2 \mathbf{I}_d$. In this section $\sigma = 0.05$.
 1182 The GLLiM model is learned with $K = 40$ to be consistent with a previous
 1183 study ([Kugler et al., 2021](#)). We check that this value is reasonable and in
 1184 particular that it cannot be significantly reduced. BIC is computed from $K = 2$
 1185 to $K = 40$. The BIC values are shown in supplementary Figure S13. The
 1186 minimum is reached for $K = 39$ but $K = 40$ provides almost the same BIC.

1187 Prior to real data inversion, performance is assessed by considering an ob-
 1188 servation simulated from the Hapke model, as explained in the supplementary
 1189 Section S4.6.2. In this experiment, ϵ is varying to observe the behavior of the
 1190 different methods (Figure S14). GLLiM-L2-ABC seems less robust, than the
 1191 other procedures, to these variations and even degrades in performance when
 1192 ϵ is too high. The two procedures based on expectations show satisfying per-
 1193 formance with globally less sharp posteriors. The addition of the posterior
 1194 log-variances does not seem to significantly change the selected samples.

1195
 1196

Reflectance measurements made in the laboratory are also generally considered by experts (see *e.g.* Pilorget et al. 2016). We focus on one observation coming from a mineral called Nontronite (see Kugler et al. 2021 for a description). 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., 2021) as likely to exhibit multiple solutions. 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 typical of real observations for which the number of possible measurements during a planet flyover is limited.

Figure 2 provides the posterior marginals for the Nontronite, obtained by setting ϵ to the 0.1% quantile of the distances. Two solutions can be deduced. Parameters ω and c show unimodal posterior distributions, while $\bar{\theta}$ distribution exhibits two modes. For b , the GLLiM-MW2-ABC sample shows a second smaller mode around 0.5 but this mode is not maintained when ϵ is set to a lower quantile (see Figure S15 in supplementary Section S4.6.3). We therefore consider that the multiplicity comes mainly from $\bar{\theta}$. In the absence of ground truth, it is difficult to fully validate the estimations. However a simple inspection consists of checking the reconstructed signals. The top-right plot in Figure 2 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 $\bar{\theta}$. The proximity of the reconstructions confirms the existence of multiple solutions and thus the relevance of a multimodal posterior. One solution can be selected by choosing the parameters that provides the best reconstruction. 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 $\bar{\theta}$ in the other solution is less physically interpretable. Note that for simplicity, we have used a uniform prior on θ but for a more meaningful study in planetary science, information on the parameters plausible values could be incorporated directly in the prior.

6.6 Computation times

The simulations ran on a laptop with 8 cores at 2.4 Ghz. Supplementary Table S3 recalls the settings and shows the computation times for the main experiments. For each experiment, the time is divided into several parts depending on the procedure. When GLLiM is used, we report the time to compute BIC from $K = 2$ to some K_{max} value, the time for learning GLLiM with the selected K value, the time to compute distances and the time for the ABC procedure per se, which consists either of rejection ABC or SMC-ABC. In the latter case, the distances computation is included in the ABC time. The compared procedures use different R packages. The computing times are therefore not fully comparable. However the overall conclusions are quite clear. The semi-automatic approach as implemented in the **abctools** package is much

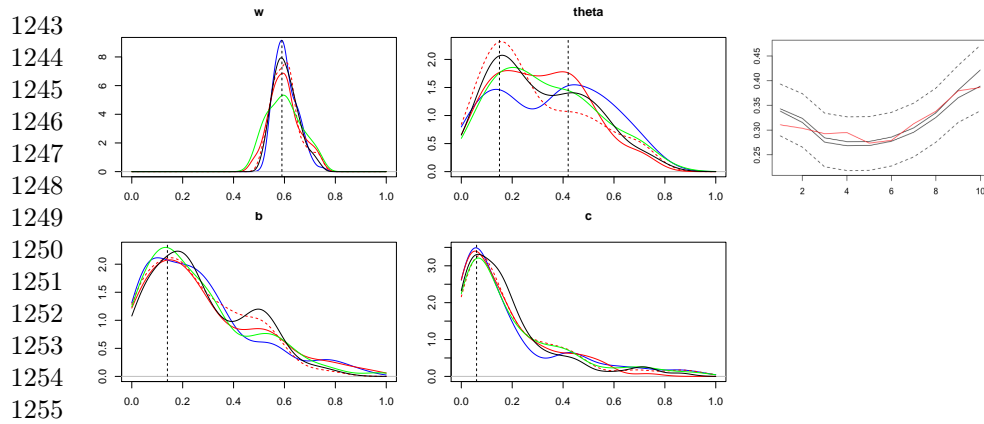


Figure 2 Real observation inversion using the Hapke model. Posterior margins for ω , θ , b and c with GLLiM-E-ABC (red), GLLiM-EV-ABC (dotted red), semi-automatic ABC (green), GLLiM-L2-ABC (blue) and GLLiM-MW2-ABC (black). The threshold ϵ is set to the 0.1% quantile (100 selected values). The vertical lines indicate the values $(\omega, \bar{\theta}, b, c) = (0.59, 0.15, 0.14, 0.06)$ and $(0.59, 0.42, 0.14, 0.06)$. The corresponding signal reconstructions (black lines) are shown in the top-right plot with the observed signal in red. The dashed lines correspond to the addition/subtraction of a standard deviation of 0.05 around the reconstructions.

faster than any other tested procedures. When dimensions of both observations and parameters are moderate and posterior distributions are likely to be unimodal, semi-automatic ABC is the most efficient choice. In contrast, GLLiM-based approaches are much more costly, especially if we include the time spent in selecting K via BIC. SMC-ABC is in general more efficient than rejection ABC even when the number of simulations is similar (see the MA(2) case). We suspect this is due to a better implementation and memory usage in the **winference** package compared to our code. The GLLiM implementation could certainly be improved but would remain based on an EM algorithm intrinsically slower. When EM is not used, as in the very special case of the normal location model, GLLiM-D-SMC procedures are actually much faster (1 to 2 minutes) *vs.* 50 minutes for WABC, which is blind to the parametric structure of the model. GLLiM-based procedures also show quite different timings depending on the experiments, ranging from a few minutes to several hours. This is due to the different GLLiM implementations (*e.g.* GLLiM-iid *vs.* standard GLLiM) and learning sets sizes and dimensions. The number of components K has also an impact on the cost of each GLLiM iteration and reflect the model complexity. For example, the Bivariate Beta model is learned with $K = 100$ in about 11 hours, which is an extreme case. We suspect this is due to the difficulty in fitting such a model. More iterations are needed for EM to converge and each iteration has a higher cost. For comparison learning GLLiM on the 14 quantiles summaries and $K = 40$ takes about 10 minutes. Reversely, the cost of computing L_2 or MW_2 distances may vary surprisingly for models of similar dimensions. The Wasserstein distance cost increases with the dimension and the number of components in the mixture. In practice, we

propose to accelerate this computation by neglecting components with too low weights. This can be quite efficient in the unimodal posterior case (1 minute 3 seconds for the MW_2 distances in the MA(2) example), while in the multiple hyperboloid example (4 hours 18 minutes for the MW_2 distances), most mixtures contain 8 components, one for each "branch", and cannot be reduced. We refer to supplementary Section S5 for more detailed comments.

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 not only could compute approximate posterior moments of higher orders as summary statistics but, more generally, approximate full posterior distributions. 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. Consequently, an interesting feature of our approach is that, with our adaptation of the original GLLiM model to *i.i.d.* data, it can be seen as an alternative to both summary-based and discrepancy-based procedures.

To handle distributions as functional summary statistics, our procedure required appropriate distances. We investigated an L_2 and a Wassertein-based distance (MW_2). The two distances often performed similarly but poor results have been observed with L_2 that would require further investigations. The MW_2 distance appeared to be more robust. 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.

Another interesting perspective would be to investigate the use of GLLiM in the context of synthetic likelihood (SL) approaches. When used in a Bayesian framework, SL techniques can be viewed as alternatives to ABC in which the intractable likelihood is replaced by an estimator of the likelihood ([Price et al., 2018](#)). Since the seminal work of [Wood \(2010\)](#), several estimators have been proposed (e.g. [Ong et al., 2018](#); [An et al., 2019, 2020](#); [Frazier and Drovandi, 2021](#)), often derived from auxiliary models ([Drovandi et al., 2015](#)). In the ABC framework of this paper, GLLiM was used to provide approximate posteriors

1335 but these posteriors are themselves coming from approximate likelihoods that
1336 could lead to new SL procedures.

1337 Lastly, in principle, any other method that is able to provide approximate
1338 surrogate posteriors could be used in place of GLLiM to produce the functional
1339 summaries. Besides the family of mixture of experts models which are similar
1340 to GLLiM, mixture density networks (Bishop, 1994) or normalizing flows (Dinh
1341 et al., 2015; Kobyzev et al., 2020; Kruse et al., 2021) are potential candidates.
1342 These neural networks have already been used in likelihood-free inference to
1343 directly approximate likelihoods or posteriors. The corresponding approaches
1344 are related to Sequential Neural Posterior Estimation (SNPE) and are different
1345 from our approach in that the approximate posteriors are not used to compute
1346 distances in a subsequent ABC scheme. SNPE is a strategy for reducing the
1347 number of simulations needed by conditional neural density estimation and is
1348 closer in spirit to SMC-ABC. These methods include SNPE-A (Papamakarios
1349 and Murray, 2016), SNPE-B (Lueckmann et al., 2017), SNPE-C or AFT
1350 (Greenberg et al., 2019). However, these methods do not all scale well with the
1351 dimension. Examples of Papamakarios and Murray (2016) are of dimension at
1352 most 10, while SNPE-C is used successfully on Lokta-Volterra time series of
1353 length 150. Overall, it is not clear whether the gain/compromise in flexibili-
1354 ty/tractability would be so much higher than with Gaussian mixtures learned
1355 with GLLiM, all the more so as GLLiM estimation could also be refined in a
1356 similar sequential learning way. A full and fair comparison would require much
1357 more work as these methods have all their own features. To the best of our
1358 knowledge, other common neural networks, like most regression techniques,
1359 would not be appropriate as they only focus on point-wise predictions.

1360

1361 **Acknowledgements.**

1362 The authors are grateful to reviewers and editors for their time and comments
1363 on this work, which have helped us in producing a much improved manuscript.
1364 FF would like to thank Guillaume Kon Kam King for an initial discussion
1365 on semi-automatic ABC, which inspired this work, Benoit Kugler and Sylvain
1366 Douté for providing the simulations for the planetary science example and for
1367 helpful discussions on the Hapke model.

1368

1369 **References**

1370

- 1371 Akesson, M., Singh, P., Wrede, F., and Hellander, A. (2021). Convolu-
1372 tional Neural Networks as Summary Statistics for Approximate Bayesian
1373 Computation. *IEEE/ACM Transactions on Computational Biology and*
1374 *Bioinformatics*.
- 1375 An, Z., Nott, D. J., and Drovandi, C. (2020). Robust Bayesian synthetic
1376 likelihood via a semi-parametric approach. *Statistics and Computing*,
1377 30(3):543–557.
- 1378 An, Z., South, L. F., Nott, D. J., and Drovandi, C. C. (2019). Accelerat-
1379 ing Bayesian Synthetic Likelihood With the Graphical Lasso. *Journal of*
1380

<i>Computational and Graphical Statistics</i> , 28(2):471–475. Publisher: Taylor & Francis.	1381 1382
Arridge, S., Maass, P., Öktem, O., and Schönlieb, C.-B. (2019). Solving inverse problems using data-driven models. <i>Acta Numerica</i> , 28:1–174.	1383 1384
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. <i>Journal of Geophysical Research: Planets</i> , 114(E6).	1385 1386 1387 1388
Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. <i>Journal of the Royal Statistical Society: Series B (Statistical Methodology)</i> , 81:235–269.	1389 1390 1391
Bishop, C. M. (1994). Mixture density networks. Technical report, Aston University, Birmingham.	1392 1393
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. <i>Statistical Science</i> , 28(2):189–208.	1394 1395 1396
Boux, F., Forbes, F., Arbel, J., Lemasson, B., and Barbier, E. L. (2021). Bayesian Inverse Regression for Vascular Magnetic Resonance Fingerprinting. <i>IEEE Trans. Medical Imaging</i> , 40(7):1827–1837.	1397 1398 1399
Buchholz, A. and Chopin, N. (2019). Improving Approximate Bayesian Computation via Quasi-Monte Carlo. <i>Journal of Computational and Graphical Statistics</i> , 28(1):205–219.	1400 1401 1402
Chen, Y., Georgiou, T. T., and Tannenbaum, A. (2019). Optimal Transport for Gaussian Mixture Models. <i>IEEE Access</i> , 7:6269–6278.	1403 1404
Chen, Y., Zhang, D., Gutmann, M., Courville, A., and Zhu, Z. (2021). Neural Approximate Sufficient Statistics for Implicit Models. In <i>ICLR2021 spotlight</i> .	1405 1406 1407
Cook, R. D. and Forzani, L. (2019). Partial least squares prediction in high-dimensional regression. <i>The Annals of Statistics</i> , 47(2):884–908.	1408 1409
Crackel, R. and Flegal, J. (2017). Bayesian inference for a flexible class of bivariate beta distributions. <i>Journal of Statistical Computation and Simulation</i> , 87:295–312.	1410 1411 1412
Csillery, K., Francois, O., and Blum, M. (2012). abc: an R package for approximate Bayesian computation (ABC). <i>Methods in Ecology and Evolution</i> .	1413 1414
Del Moral, P., Doucet, A., and Jasra, A. (2012). An Adaptive Sequential Monte Carlo Method for Approximate Bayesian Computation. <i>Statistics and Computing</i> , 22(5):1009–1020.	1415 1416 1417
Deleforge, A., Forbes, F., Ba, S., and Horaud, R. (2015a). Hyper-Spectral Image Analysis with Partially-Latent Regression and Spatial Markov Dependencies. <i>IEEE Journal of Selected Topics in Signal Processing</i> , 9(6):1037–1048.	1418 1419 1420 1421
Deleforge, A., Forbes, F., and Horaud, R. (2015b). High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables. <i>Statistics and Computing</i> , 25(5):893–911.	1422 1423 1424 1425 1426

- 1427 Delon, J. and Desolneux, A. (2020). A Wasserstein-type distance in the space
1428 of Gaussian Mixture Models. *SIAM Journal on Imaging Sciences*.
- 1429 Dinh, L., Krueger, D., and Bengio, Y. (2015). NICE: non-linear independent
1430 components estimation. In Bengio, Y. and LeCun, Y., editors, *3rd Inter-*
1431 *national Conference on Learning Representations, ICLR 2015, San Diego,*
1432 *CA, USA, May 7-9, 2015, Workshop Track Proceedings*.
- 1433 Drovandi, C., Pettitt, T., and Lee, A. (2015). Bayesian indirect inference using
1434 a parametric auxiliary model. *Statistical Science*, 30(1):72–95.
- 1435 Drovandi, C. C. and Pettitt, A. N. (2011). Likelihood-free Bayesian estimation
1436 of multivariate quantile distributions. *Computational Statistics and Data*
1437 *Analysis*, 55:2541–2556.
- 1438 Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for
1439 approximate Bayesian computation: semi-automatic approximate Bayesian
1440 computation. *Journal of the Royal Statistical Society: Series B (Statistical*
1441 *Methodology)*, 74(3):419–474.
- 1442 Fernando, J., Schmidt, F., and Douté, S. (2016). Martian surface microtexture
1443 from orbital CRISM multi-angular observations: A new perspective for the
1444 characterization of the geological processes. *Planetary and Space Science*,
1445 128:30–51.
- 1446 Frazier, D. T. and Drovandi, C. (2021). Robust Approximate Bayesian Infer-
1447 ence With Synthetic Likelihood. *Journal of Computational and Graphical*
1448 *Statistics*, pages 1–19.
- 1449 Greenberg, D., Nonnenmacher, M., and Macke, J. (2019). Automatic posterior
1450 transformation for likelihood-free inference. In *International Conference on*
1451 *Machine Learning*, pages 2404–2414. PMLR.
- 1452 Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2018). Likelihood-free
1453 inference via classification. *Statistics and Computing*, 28:411–425.
- 1454 Hovorka, R., Canonico, V., Chassin, L. J., Haueter, U., Massi-Benedetti, M.,
1455 Federici, M. O., Pieber, T. R., Schaller, H. C., Schaupp, L., Vering, T.,
1456 and Wilinska, M. E. (2004). Nonlinear model predictive control of glucose
1457 concentration in subjects with type 1 diabetes. *Physiological Measurement*,
1458 25(4):905–920.
- 1459 Ingrassia, S., Minotti, S. C., and Vittadini, G. (2012). Local Statistical Model-
1460 ing via a Cluster-Weighted Approach with Elliptical Distributions. *Journal*
1461 *of classification*, 29(3):363–401.
- 1462 Jacob, P., Bernton, E., Gerber, M., and Robert, C. P. (2020). *Winference: R*
1463 *package to perform approximate Bayesian computation with the Wasserstein*
1464 *distance*.
- 1465 Jiang, B., Wu, T.-Y., C., Z., and Wong, W. (2017). Learning summary
1466 statistics for Approximate Bayesian Computation via Deep Neural Network.
1467 *Statistica Sinica*, pages 1595–1618.
- 1468 Jiang, B., Wu, T.-Y., and Wong, W. H. (2018). Approximate Bayesian
1469 computation with Kullback-Leibler divergence as data discrepancy. In
1470 *21st International Conference on Artificial Intelligence and Statistics (AIS-*
1471 *TATS)*.
- 1472

- Kobyzev, I., Prince, S., and Brubaker, M. (2020). Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–1. 1473–1475
- Kristan, M., Leonardis, A., and Skočaj, D. (2011). Multivariate online kernel density estimation with Gaussian kernels. *Pattern Recognition*, 44(10-11):2630–2642. 1476–1478
- Kruse, J., Ardizzone, L., Rother, C., and Kothe, U. (2021). Benchmarking invertible architectures on inverse problems. *Workshop on Invertible Neural Networks and Normalizing Flows (ICML 2019)*, arXiv preprint arXiv:2101.10763. 1479–1482
- Kugler, B., Forbes, F., and Douté, S. (2021). Fast Bayesian Inversion for high dimensional inverse problems. *To appear in Statistics and Computing*, <https://hal.archives-ouvertes.fr/hal-02908364>. 1483–1485
- 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. 1486–1489
- 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. 1490–1493
- Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of American Statistical Association*, 86(414):316–327. 1494–1495
- Lueckmann, J.-M., Boelts, J., Greenberg, D. S., Gonçalves, P. J., and Macke, J. H. (2021). Benchmarking simulation-based inference. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *Proceedings of Machine Learning Research*, pages 343–351. PMLR. 1496–1499
- Lueckmann, J.-M., Gonçalves, P. J., Bassetto, G., Öcal, K., Nonnenmacher, M., and Macke, J. H. (2017). Flexible statistical inference for mechanistic models of neural dynamics. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. 1500–1505
- 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. 1506–1508
- Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computation methods. *Statistics and Computing*, 22:1167–1180. 1509–1510
- Mesejo, P., Sallet, 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. 1511–1514
- 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. 1515–1518

- 1519 Murchie, S. L., Seelos, F. P., Hash, C. D., Humm, D. C., Malaret, E., Mc-
1520 Govern, J. A., Choo, T. H., Seelos, K. D., Buczkowski, D. L., Morgan,
1521 M. F., Barnouin-Jha, O. S., Nair, H., Taylor, H. W., Patterson, G. W.,
1522 Harvel, C. A., Mustard, J. F., Arvidson, R. E., McGuire, P., Smith, M. D.,
1523 Wolff, M. J., Titus, T. N., Bibring, J.-P., and Poulet, F. (2009). Compact
1524 Reconnaissance Imaging Spectrometer for Mars investigation and data set
1525 from the Mars Reconnaissance Orbiter’s primary science phase. *Journal of*
1526 *Geophysical Research: Planets*, 114(E2):E00D07.
- 1527 Nataraj, G., Nielsen, J.-F., Scott, C., and Fessler, J. A. (2018). Dictionary-
1528 Free MRI PERK: Parameter Estimation via Regression with Kernels. *IEEE*
1529 *Trans. Med. Imaging*, 37(9):2103–2114.
- 1530 Nguyen, H. D., Arbel, J., Lu, H., and Forbes, F. (2020a). Approximate
1531 Bayesian Computation Via the Energy Statistic. *IEEE Access*, 8:131683–
1532 131698.
- 1533 Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2019). Approximation re-
1534 sults regarding the multiple-output Gaussian gated mixture of linear experts
1535 model. *Neurocomputing*.
- 1536 Nguyen, H. D., Nguyen, T., Chamroukhi, F., and McLachlan, G. J. (2021a).
1537 Approximations of conditional probability density functions in Lebesgue
1538 spaces via mixture of experts models. *Journal of Statistical Distributions*
1539 *and Applications*, 8(1):13.
- 1540 Nguyen, T., Chamroukhi, F., Nguyen, H. D., and McLachlan, G. J. (2020b).
1541 Approximation of probability density functions via location-scale finite mix-
1542 tures in Lebesgue spaces. *arXiv preprint arXiv:2008.09787*. To appear.
1543 *Communications in Statistics - Theory and Methods*.
- 1544 Nguyen, T., Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2021b). A non-
1545 asymptotic penalization criterion for model selection in mixture of experts
1546 models. *arXiv preprint arXiv:2104.02640*.
- 1547 Nguyen, T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2020c).
1548 Approximation by finite mixtures of continuous density functions that vanish
1549 at infinity. *Cogent Mathematics & Statistics*, 7(1):1750861.
- 1550 Nunes, M. A. and Prangle, D. (2015). abctools: An R package for
1551 tuning Approximate Bayesian Computation analyses. [https://cran.r-](https://cran.r-project.org/web/packages/abctools/)
1552 [project.org/web/packages/abctools/](https://cran.r-project.org/web/packages/abctools/).
- 1553 Ong, V., Nott, D., Tran, M.-N., Sisson, S., and Drovandi, C. (2018). Likelihood-
1554 free inference in high dimensions with synthetic likelihood. *Computational*
1555 *Statistics and Data Analysis*, 128.
- 1556 Papamakarios, G. and Murray, I. (2016). Fast ε -Free Inference of Simula-
1557 tion Models with Bayesian Conditional Density Estimation. In Lee, D.,
1558 Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in*
1559 *Neural Information Processing Systems*, volume 29. Curran Associates, Inc.
- 1560 Park, M., Jitkrittum, W., and Sejdinovic, D. (2016). K2-ABC: approxi-
1561 mate Bayesian computation with kernel embeddings. In *19th International*
1562 *Conference on Artificial Intelligence and Statistics (AISTATS)*.
1563
1564

- Perthame, E., Forbes, F., Deleforge, A., Devijver, E., and Gallopin, M. (2017). *xLLiM: High Dimensional Locally-Linear Mapping*. R package version 2.1. 1565–1566
- 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. 1567–1569
- 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. 1570–1571–1572
- Price, L. F., Drovandi, C. C., Lee, A., and Nott, D. J. (2018). Bayesian Synthetic Likelihood. *Journal of Computational and Graphical Statistics*, 27(1):1–11. 1573–1574–1575
- Rakhlín, A., Panchenko, D., and Mukherjee, S. (2005). Risk bounds for mixture density estimation. *ESAIM: Probability and Statistics*, 9:220–229. 1576–1577
- Rodrigues, G. S., Nott, D. J., and Sisson, S. A. (2016). Functional regression approximate Bayesian computation for Gaussian process density estimation. *Computational Statistics & Data Analysis*, 103:229–241. 1578–1579–1580
- Rubio, F. and Johansen, A. M. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, 7:1632–1654. 1581–1582–1583
- Schmidt, F. and Fernando, J. (2015). Realistic uncertainties on Hapke model parameters from photometric measurements. *Icarus*, 260:73–93. 1584–1585
- Sisson, S. A., Fan, Y., and Beaumont, M. A., editors (2019). *Handbook of Approximate Bayesian Computation*. CRC Press, Boca Raton. 1586–1587
- Soubeyrand, S., Carpentier, F., Guiton, F., and Klein, E. K. (2013). Approximate Bayesian computation with functional statistics. *Statistical Applications in Genetics and Molecular Biology*, 12(1):17–37. 1588–1589–1590
- 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. 1591–1592–1593
- 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. 1594–1595–1596–1597–1598
- Wiqvist, S., Mattei, P.-A., Picchini, U., and Frelsen, 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. 1599–1600–1601–1602–1603
- Wood, S. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104. 1604–1605–1606–1607–1608–1609–1610