Mixture of expert posterior surrogates for approximate Bayesian computation

Florence Forbes, Hien Duy Nguyen, Trungtin Nguyen, Julyan Arbel

Mixture of expert posterior surrogates for approximate Bayesian computation

Florence Forbes [1], Hien Duy Nguyen [3], TrungTin Nguyen [1] & Julyan Arbel [1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*
florence.forbes@inria.fr, trung-tin.nguyen@inria.fr, julyan.arbel@inria.fr
[3] *School of Mathematics and Physics, University of Queensland, St. Lucia, Australia.*
h.nguyen7@uq.edu.au

Résumé. Les procédures de calcul bayésien approché (ABC) reposent sur l'évaluation de l'écart entre les données simulées et les données observées. Cet écart est souvent évalué en comparant des statistiques résumées plutôt que directement les données. Le choix d'une distance et de résumés appropriés est donc une étape cruciale qui peut affecter la qualité des approximations. Dans ce travail, nous introduisons une étape d'apprentissage préliminaire dans laquelle des lois de substitution, issues d'un modèle de mélange d'experts, sont construites pour approximer les lois a posteriori visées. Ces lois a posteriori de substitution sont ensuite utilisées à la place des statistiques résumées et comparées à l'aide de métriques entre distributions. On montre que la quasi-loi a posteriori résultante converge vers la vraie loi a posteriori, sous des conditions standard. Des expériences montrent que notre approche est particulièrement performante lorsque la loi a posteriori est multimodale.

Mots-clés. Approximate Bayesian computation, Mélanges d'experts, statistiques résumées, distance de Wasserstein, modèles de substitution, lois a posteriori mulitmodales.

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 expert or prior knowledge, which seldom occurs, they have to be chosen and this 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 using a specific instance of a Mixture of Experts model. These surrogate posteriors are then used in place of summary statistics and compared using metrics between distributions in place of data discrepancies. The resulting ABC quasi-posterior distribution is shown to converge to the true one, under standard conditions. Experiments show that our approach is particularly useful when the posterior is multimodal.

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

1 Introduction

Approximate Bayesian computation (ABC) appears as a natural candidate for addressing estimation problems where there is a lack of availability or tractability of the likelihood, but when the data generating process is available as a tractable simulation procedure. The fundamental idea of ABC is to generate parameter proposals $\boldsymbol{\theta}$ in a parameter space Θ using a prior distribution $\pi(\boldsymbol{\theta})$ and accept a proposal if the simulated data $\mathbf{z}$ for that proposal is similar to the observed data $\mathbf{y}$, where $\mathbf{z}, \mathbf{y} \in \mathcal{Y}$. This similarity is usually measured using a distance or discriminative measure D, whereby a simulated sample $\mathbf{z}$ is retained if $D(\mathbf{z}, \mathbf{y})$ is smaller than a given threshold ϵ. In this simple form, the procedure is generally referred to as rejection ABC. Other variants are possible and often recommended, for instance using MCMC or sequential procedures. We will focus on the rejection version as all developments can be easily adapted to more sophisticated variants.

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

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) \propto \pi(\boldsymbol{\theta}) \int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}} \, f_{\boldsymbol{\theta}}(\mathbf{z}) \, d\mathbf{z} \, . \tag{1.1}$$

More specifically, the similarity between $\mathbf{z}$ and $\mathbf{y}$ is 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 (1.1) 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 (Fearnhead and Prangle, 2012). 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.

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 combinations of posterior moments, not restricted to the posterior means. For this purpose, the Gaussian Locally Linear Mapping (GLLiM) method (Deleforge et al., 2015) appears as a good candidate regression model, with properties that balance between computationally expensive neural networks and simple standard regression techniques. GLLiM provides, at

low cost, a parametric estimation of the true posterior distributions. Using a learning set of parameter and observation pairs, GLLiM learns a family of finite Gaussian mixtures whose parameters depend analytically on the observation to be inverted. Such models correspond to Gaussian mixture of experts whose moments can be easily computed and used as summary statistics.

Our second contribution is to propose to compare directly the full surrogate posterior distributions provided by GLLiM, without reducing them to their moments. This requires the specification of a distance to compare such distributions. The recent Mixture-Wasserstein distance, denoted throughout the text as MW_2, designed for Gaussian mixtures (Delon and Desolneux, 2020) match perfectly this need. There exist other distances between mixtures that are tractable such as the L_2 distance, which is also considered in this work.

A remarkable feature of our approach is that it can be equally applied to settings where a sample of *i.i.d.* observations is available (*e.g.* Bernton et al. 2019) and to settings where a single observation is available as a vector, a time series realization, or a data set that is reduced to a vector of summary statistics (*e.g.* Fearnhead and Prangle 2012).

Regarding the approach's theoretical properties, we provide two results. In the first result, the true posterior is used to compare samples $\mathbf{y}$ and $\mathbf{z}$. In the second result, 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. More details can be found in Forbes et al. (2021).

2 Extended semi-automatic ABC

A learning set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ is built from the joint distribution that results from the prior $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ and the likelihood $f_{\boldsymbol{\theta}}$, where $[N] = \{1, \ldots, N\}$. The idea is to capture the relationship between $\boldsymbol{\theta}$ and $\mathbf{y}$ with a joint probability model with computationally inexpensive and straightforward conditional distributions and moments. For the choice of the model to fit to $\mathcal{D}_N$, we propose to use the so-called GLLiM model (Deleforge et al., 2015) for its ability to capture non-linear relationships in a tractable manner, based on flexible mixtures of Gaussian distributions. GLLiM provides, for each observed $\mathbf{y}$, a full posterior probability distribution within a family of parametric Mixture of Expert models $\{p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}), \boldsymbol{\phi} \in \boldsymbol{\Phi}\}$. To model non-linear relationships, it uses a mixture of K linear models. More specifically, the expression of $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi})$ is analytical and available for all $\mathbf{y}$ with $\boldsymbol{\phi}$ being independent of $\mathbf{y}$:

$$p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}) = \sum_{k=1}^{K} \eta_k(\mathbf{y})\, \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{A}_k\mathbf{y} + \boldsymbol{b}_k, \boldsymbol{\Sigma}_k), \tag{2.1}$$

where $\mathcal{N}(\,\cdot\,; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian pdf with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and $\eta_k(\mathbf{y}) = \pi_k \mathcal{N}(\mathbf{y}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k) / \sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{y}; \boldsymbol{c}_j, \boldsymbol{\Gamma}_j)$. This distribution involves parameters:

$\boldsymbol{\phi} = \{\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k, \boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$. The parameter $\boldsymbol{\phi}$ can be estimated using an Expectation-Maximization (EM) algorithm (Deleforge et al., 2015). Fitting a GLLiM model to $\mathcal{D}_N$ therefore results in a set of parametric distributions $\{p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N}),\ \mathbf{y} \in \mathcal{Y}\}$, which 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 $\mathbf{y}$.

Instead of comparing simulated $\mathbf{z}$'s to the observed $\mathbf{y}$, or comparing their summary statistics, we propose to compare $p_G(\boldsymbol{\theta} \mid \mathbf{z}; \boldsymbol{\phi}^*_{K,N})$'s and $p_G(\boldsymbol{\theta} \mid \mathbf{y}; \boldsymbol{\phi}^*_{K,N})$, as given by (2.1). We then derive two procedures, referred to as GLLiM-MW2-ABC and GLLiM-L2-ABC in Algorithm 1, respectively.

Algorithm 1 GLLiM-ABC algorithms – Vector and functional variants

1: **Inverse operator learning.** Apply GLLiM on a training set $\mathcal{D}_N = \{(\boldsymbol{\theta}_n, \mathbf{y}_n), n \in [N]\}$ to estimate, for any $\mathbf{z} \in \mathcal{Y}$, the K-Gaussian mixture $p_G(\boldsymbol{\theta} \mid \mathbf{z}; \boldsymbol{\phi}^*_{K,N})$ in (2.1) as a first approximation of the true posterior $\pi(\boldsymbol{\theta} \mid \mathbf{z})$, where $\boldsymbol{\phi}^*_{K,N}$ does not depend on $\mathbf{z}$.

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

Vector summary statistics.

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

GLLiM-EV-ABC: Compute both $s_1(\mathbf{z}_m)$ and $s_2(\mathbf{z}_m) = \mathrm{Var}_G[\boldsymbol{\theta} \mid \mathbf{z}_m; \boldsymbol{\phi}^*_{K,N}]$.

In both cases, compute standard distances between summary statistics.

Functional summary statistics.

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

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

3: **Sample selection.** Select $\boldsymbol{\theta}_m$ values that lead to distances under an ϵ threshold (rejection ABC) or apply an ABC procedure that can handle distances, directly.

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

3 Theoretical properties

One important question is the proximity of the resulting so-called ABC quasi-posterior to the true posterior. We assume a fixed given observed $\mathbf{y}$ and the dependence on $\mathbf{y}$ is omitted from the notation. Let us first recall the standard form of the ABC quasi-posterior, omitting summary statistics from the notation:

$$\pi_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}}\ \pi(\boldsymbol{\theta} \mid \mathbf{z})\ \pi(\mathbf{z})\ d\mathbf{z}}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\mathbf{y},\mathbf{z}) \leq \epsilon\}}\ \pi(\mathbf{z})\ d\mathbf{z}}\ . \tag{3.1}$$

If D is a distance and $D(\mathbf{y}, \mathbf{z})$ is continuous in $\mathbf{z}$, the ABC posterior in (3.1) can be shown to have the desirable property of converging to the true posterior when ϵ tends to 0. The proof is based on the fact that when ϵ tends to 0, due to the property of the distance D, the set $\{\mathbf{z} \in \mathcal{Y} : D(\mathbf{y}, \mathbf{z}) \leq \epsilon\}$, defining the indicator function in (3.1), tends to the singleton $\{\mathbf{y}\}$ so that consequently $\mathbf{z}$ in the likelihood can be replaced by the observed $\mathbf{y}$, which then leads to an ABC quasi-posterior proportional to $\pi(\boldsymbol{\theta}) f_{\boldsymbol{\theta}}(\mathbf{y})$ and therefore to the true posterior as desired. It is interesting to note that this proof is based on working on the term under the integral only and is using the equality, at the limit, of $\mathbf{z}$ to $\mathbf{y}$, which is actually a stronger assumption than necessary for the result to hold.

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

$$q_\epsilon(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}), \pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\boldsymbol{\theta} \mid \mathbf{z}) \, \pi(\mathbf{z}) \, d\mathbf{z}}{\int_{\mathcal{Y}} \mathbf{1}_{\{D(\pi(\cdot|\mathbf{y}), \pi(\cdot|\mathbf{z})) \leq \epsilon\}} \pi(\mathbf{z}) \, d\mathbf{z}} \,. \tag{3.2}$$

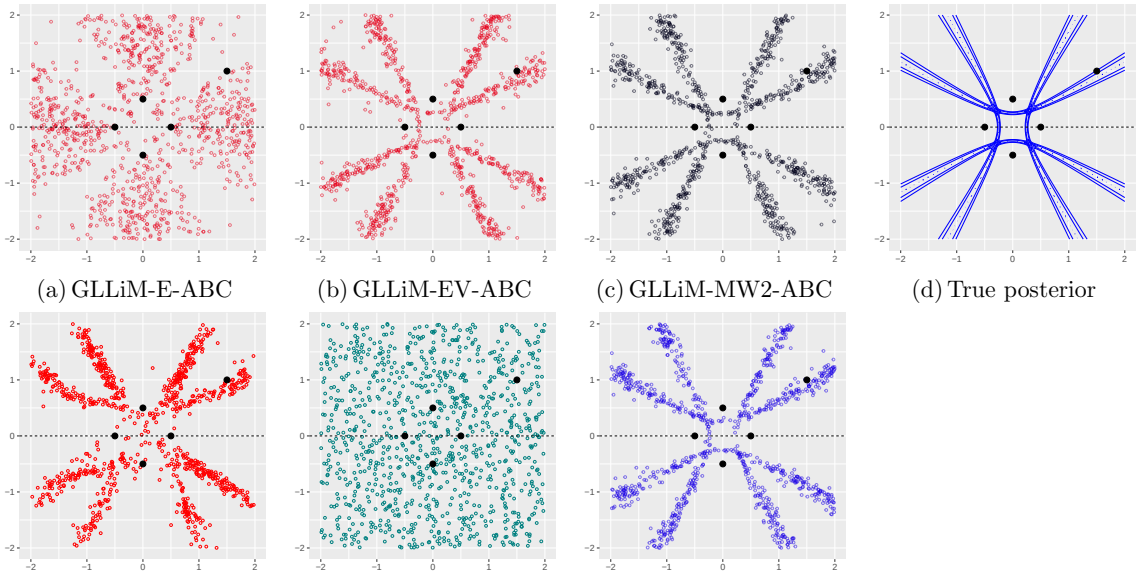
The following Theorem 3.1, proved in Forbes et al. (2021), shows that $q_\epsilon(\cdot \mid \mathbf{y})$ converges to $\pi(\cdot \mid \mathbf{y})$ in total variation, for fixed $\mathbf{y}$. For $\epsilon > 0$, $A_\epsilon = \{\mathbf{z} \in \mathcal{Y} : D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z})) \leq \epsilon\}$.

Theorem 3.1. *If $\pi(\boldsymbol{\theta} \mid \cdot)$ is continuous for all $\boldsymbol{\theta} \in \Theta$, and $\sup_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} \mid \mathbf{y}) < \infty$; there exists a $\gamma > 0$ such that $\sup_{\boldsymbol{\theta} \in \Theta} \sup_{\mathbf{z} \in A_\gamma} \pi(\boldsymbol{\theta} \mid \mathbf{z}) < \infty$; $D(\cdot, \cdot) : \Pi \times \Pi \to \mathbb{R}_+$ is a metric on the functional class $\Pi = \{\pi(\cdot \mid \mathbf{y}) : \mathbf{y} \in \mathcal{Y}\}$; and $D(\pi(\cdot \mid \mathbf{y}), \pi(\cdot \mid \mathbf{z}))$ is continuous, with respect to $\mathbf{z}$. Then, $q_\epsilon(\cdot \mid \mathbf{y})$ in (3.2) converges in total variation to $\pi(\cdot \mid \mathbf{y})$, for fixed $\mathbf{y}$, as $\epsilon \to 0$.*

In most ABC settings, based on data discrepancy or summary statistics, the above consideration and result are not useful because the true posterior is unknown by construction and cannot be used to compare samples. However this principle becomes useful in our setting, which is based on surrogate posteriors. While the previous result can be seen as an oracle of sorts, it is more interesting in practice to investigate whether a similar result holds when using surrogate posteriors in the ABC likelihood. This is the goal of Theorem 2 in Forbes et al. (2021) which shows the convergence of the ABC quasi-posterior for a restricted class of target distributions and surrogate posteriors learned as mixtures.

4 Illustration

Numerical experiments, available in Forbes et al. (2021), show the versatility of our approach, applicable on both *i.i.d.* samples and single observation settings. The example below shows that it is particularly useful, when the posterior is multimodal. The object of interest is an unknown parameter $\boldsymbol{\theta} = (x, y)$ that can be interpreted as a source location in a 2D scene. To create a multimodal posterior, we consider a likelihood detailed in Forbes et al. (2021). The true posterior is shown in Figure 1 (d).



(a) GLLiM-E-ABC (b) GLLiM-EV-ABC (c) GLLiM-MW2-ABC (d) True posterior

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

Figure 1: GLLiM is learned with $K = 38$ and $N = 10^5$ while ABC is run usiing $M = 10^6$ simulations for (a,b,f,h) and $M = 10^5$ for (c,g). (a) GLLiM-E-ABC, (b) GLLiM-EV-ABC, (c) GLLiM-MW2-ABC, (d) contours of the true posterior, (e) approximate GLLiM posterior for the observed data, (f) semi-automatic ABC and (g) GLLiM-L2-ABC.

References

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81:235–269. 2, 3

Deleforge, A., Forbes, F., and Horaud, R. (2015). High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables. *Statistics and Computing*, 25(5):893–911. 2, 3, 4

Delon, J. and Desolneux, A. (2020). A Wasserstein-type distance in the space of Gaussian Mixture Models. *SIAM Journal on Imaging Sciences*. 3

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474. 2, 3

Forbes, F., Nguyen, H. D., Nguyen, T. T., and Arbel, J. (2021). Approximate Bayesian computation with surrogate posteriors. *Preprint. hal-03139256*. 3, 5