Bayesian Likelihood Free Inference using Mixtures of Experts

Florence Forbes, Hien Duy Nguyen, Trungtin Nguyen

Bayesian Likelihood Free Inference using Mixtures of Experts

Florence Forbes, Hien Duy Nguyen, TrungTin Nguyen

***Abstract*—We extend Bayesian Synthetic Likelihood (BSL) methods to non-Gaussian approximations of the likelihood function. In this setting, we introduce Mixtures of Experts (MoEs), a class of neural network models, as surrogate likelihoods that exhibit desirable approximation theoretic properties. Moreover, MoEs can be estimated using Expectation–Maximization algorithm-based approaches, such as the Gaussian Locally Linear Mapping model estimators that we implement. Further, we provide theoretical evidence towards the ability of our procedure to estimate and approximate a wide range of likelihood functions. Through simulations, we demonstrate the superiority of our approach over existing BSL variants in terms of both posterior approximation accuracy and computational efficiency.**

***Index Terms*—Likelihood Free Inference, Bayesian Synthetic Likelihood, Mixture of Experts, Gaussian Locally Linear Mapping**

I. Introduction

Likelihood-free inference, or simulation-based inference, entails estimation of parameters $\boldsymbol{\theta}$ of a stochastic model without a feasible likelihood function. However, we assume the ability to simulate observations given known parameters. This study takes a Bayesian stance, where we aim to estimate posterior parameter distributions.

In this work, we concentrate on methods that utilize surrogate parametric models in place of an intractable likelihoods. Specifically, we study the class of methods generally called *Synthetic likelihood* (SL) procedures. These methods provide estimates of likelihood functions that are then used as inputs for a sampling procedure, such as a Markov chain Monte Carlo (MCMC) scheme, to estimate the posterior distribution. Bayesian Synthetic Likelihood (BSL) approaches [23] have, in the past, been investigated as Bayesian extensions of the SL approach of [31]. SL is typically characterized by a Gaussian assumption, while more general formulations are studied under the parametric Bayesian indirect likelihood (pBIL) framework, which includes a number of variants [5]. For comparisons with our approaches, we focus on the variants implemented in the BSL package in R [2].

The typical approach of BSL methods is to approximate the intractable likelihood by a multivariate Gaussian distribution whose mean and covariance parameters depend on $\boldsymbol{\theta}$ and are estimated pointwise for each value of $\boldsymbol{\theta}$ via the empirical mean and covariance estimators of a sample of m independent and identically distributed (*i.i.d.*) summary statistics, simulated from the underlying likelihood, respectively (cf. [23], [31]).

For good performance, the number of simulations m should not be too small, and evidence suggests that, in practice, the ideal m increases as the dimension of the summary statistics grows [23]. Various approaches have been explored to decrease the required number of simulations, such as sparse techniques [3] and shrinkage techniques [22], [24], which aim to diminish the number of parameters necessary for estimating the covariance matrix. Additionally, the *uBSL* approach of [23] consider unbiased estimation of the normal density functional rather than its mean and covariance parameters. A Semi-parametric variant, *semiBSL*, has also been suggested to provide robustness when likelihoods are non-Gaussian [1].

In [7], two other variants are considered, referred to as *missBSL*. They aim to estimate the Gaussian synthetic likelihood in a more robust manner, to account for incompatibilities between model and summary choice, *i.e.*, model misspecification. The first approach, denoted as *missBSLmean*, augments the mean of the simulated summaries with additional free parameters, while the second approach, *missBSLvar*, augments the variance with free parameters, instead.

For all described BSL alternatives, above, an MCMC scheme is required to carry out the posterior inference. Therefore, if I evaluations of the likelihood are needed in the subsequent MCMC algorithm, it is necessary to simulate I values of $\boldsymbol{\theta}$ according to the prior and then simulate $I \times m$ values of observations $\mathbf{y}$ due to the pointwise construction of the SL estimators. For large I and m, this can be overly costly. The solution we investigate is based on Mixture of Experts (MoE), a class of neural network models [15], [33], using the so-called Gaussian Locally Linear Mapping model (GLLiM; [4], [32]) estimator. Our approach has the advantage to both reduce the number of simulations m and depart from Gaussianity assumptions. Additionally, it allows us to exploit recent approximation and estimation theoretic results regarding MoEs [16]–[20] to establish desirable theoretical results. These results fill a gap in the BSL literature, where there is a lack of theory based on mild and easily checkable assumptions that allow for guarantees in the relationship between the estimated BSL posterior and the target posterior measures.

II. Bayesian synthetic likelihood

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space. We observe data $\mathbf{X}_n = (\boldsymbol{X}_i)_{i\in[n]}$, where $[n] = \{1, \ldots, n\}$ and $\boldsymbol{X}_i : \Omega \to \mathbb{X}$ is a random variable taking value in the measurable space $(\mathbb{X}, \mathfrak{X})$, for each $i \in [n]$. We can thus endow $\mathbf{X}_n$ with the push-forward probability space $(\mathbb{X}^n, \mathfrak{X}^{\otimes n}, \mathbb{P}_n)$. Further define the parameter space $(\mathbb{T}, \mathfrak{T})$, with typical element $\boldsymbol{\theta}$, and equip it with the prior measure $\Pi : \mathfrak{T} \to [0, 1]$. In classical Bayesian inference (see *e.g.*, [25] and [29]), one assumes that $(\boldsymbol{\theta}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ has joint measure $(\mathbb{T} \times \mathbb{X}^n, \mathfrak{T} \otimes \mathfrak{X}^{\otimes n}, \mathbb{Q}_n)$, where $\mathbb{Q}_n \ll \lambda$

for some dominating measure λ (*e.g.* Lebesgue or counting measures), where, for each $\mathbb{A} \in \mathfrak{T} \otimes \mathfrak{X}^{\otimes n}$:

$$\mathbb{Q}_n(\mathbb{A}) = \int_{\mathbb{A}} f_n(\mathbf{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\lambda(\mathrm{d}\boldsymbol{\theta} \times \mathrm{d}\mathbf{x}_n).$$

We typically call $f_n : \mathbb{X}^n \times \mathbb{T} \to \mathbb{R}_{\geq 0}$ the likelihood function, with the property that $\int_{\mathbb{X}^n} f_n(\mathbf{x}_n|\boldsymbol{\theta})\lambda(\mathrm{d}\mathbf{x}_n) = 1$, for each $\boldsymbol{\theta} \in \mathbb{T}$, and where $\pi : \mathbb{T} \to \mathbb{R}_{\geq 0}$ is the density of the prior measure Π, with respect to λ. The target of Bayesian inference is to either provide an expression for the posterior measure $\Pi(\cdot|\mathbf{x}_n) : \mathfrak{T} \to [0, 1]$, characterized by integration with respect to the posterior density $\pi(\boldsymbol{\theta}|\mathbf{x}_n) \propto f_n(\mathbf{x}_n|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ or to construct Monte Carlo estimators for integrals with respect to $\Pi(\cdot|\mathbf{x}_n)$.

A. BSL original and variants

Most Bayesian approaches require a closed-form expression for f_n and cannot be used in the likelihood-free setting. In the BSL setting, f_n can be intractable but we assume that we can simulate *i.i.d.* samples: $\mathbf{Y}_N^m = (\boldsymbol{\theta}_j, \mathbf{X}_{n,j}^1, \ldots, \mathbf{X}_{n,j}^m)_{j\in[N]}$, where $\boldsymbol{\theta}_j$ has prior measure Π and $\mathbf{X}_{n,j}^k|\boldsymbol{\theta}_j$ has measure $f_n(\cdot|\boldsymbol{\theta})\mathrm{d}\lambda$, for $k \in [m]$. We then use these data to estimate some tractable replacement for the likelihood function: $g_n(\boldsymbol{\eta}(\cdot)|\boldsymbol{\theta}) : \mathbb{X}^n \to \mathbb{R}_{\geq 0}$, with $\int_{\mathbb{R}^d} g_n(\boldsymbol{\eta}|\boldsymbol{\theta})\lambda(\mathrm{d}\boldsymbol{\eta}) = 1$, for each $\boldsymbol{\theta} \in \mathbb{T}$ and a summary statistic $\boldsymbol{\eta} : \mathbb{X}^n \to \mathbb{R}^d$. In particular, the approach of [23] suggests to use simulations $\mathbf{Y}_N^m$ to estimate the likelihood replacement

$$g_{n,m}(\boldsymbol{\eta}(\mathbf{x}_n)|\boldsymbol{\theta}) = \int_{\mathbb{R}^{d\times m}} \mathcal{N}_d(\boldsymbol{\eta}(\mathbf{x}_n); \boldsymbol{\mu}((\mathbf{x}_n^k)_{k\in[m]}), \boldsymbol{\Sigma}((\mathbf{x}_n^k)_{k\in[m]})) \times \prod_{k=1}^m f_n(\mathbf{x}_n^k|\boldsymbol{\theta})\lambda(\mathrm{d}(\mathbf{x}_n^k)_{k\in[m]}), \quad (1)$$

where

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}((\mathbf{x}_n^k)_{k\in[m]}) = \frac{1}{m}\sum_{k=1}^m \boldsymbol{\eta}(\mathbf{x}_n^k), \quad (2)$$

$$\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}((\mathbf{x}_n^k)_{k\in[m]}) = \frac{1}{m}\sum_{k=1}^m \boldsymbol{\eta}(\mathbf{x}_n^k)\boldsymbol{\eta}^\top(\mathbf{x}_n^k) - \frac{1}{m}\boldsymbol{\mu}((\mathbf{x}_n^k)_{k\in[m]})\boldsymbol{\mu}^\top((\mathbf{x}_n^k)_{k\in[m]}) \quad (3)$$

and $\mathcal{N}_d(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the d-dimensional normal density function, with mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and positive definite covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$.

Variations on this construction have been proposed, for example, by [1] and [7]. In [1] the authors propose to replace $\mathcal{N}_d(\boldsymbol{\eta}(\mathbf{x}_n); \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, in (1), by a copula transformation of a marginal kernel density estimator, leading to the *semiBSL*. In [7], the authors introduce a prior on an additional free parameter that improves robustness to misspecification of choice of summary statistic $\boldsymbol{\eta}$, leading to the *missBSL* approach. Further refinements have subsequently been considered by [8], where the covariance estimator (3) is replaced by more general classes of covariance estimators.

B. Theoretical insights of BSL

It is noteworthy that a number of theoretical results have been proved with respect to the described BSL algorithms. Firstly, for fixed $\mathbf{x}_n$, [5] proved a weak consistency result regarding the approximate posterior measure defined by density

$$g_m(\boldsymbol{\theta}|\mathbf{x}_n) \propto g_{n,m}(\boldsymbol{\eta}(\mathbf{x}_n)|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

to a limiting posterior measure defined by

$$g(\boldsymbol{\theta}|\mathbf{x}_n) \propto g_n(\boldsymbol{\eta}(\mathbf{x}_n)|\boldsymbol{\theta})\pi(\boldsymbol{\theta}),$$

where, $g_n(\boldsymbol{\eta}(\mathbf{x}_n)|\boldsymbol{\theta}) = \mathcal{N}_d(\boldsymbol{\eta}(\mathbf{x}_n); \boldsymbol{\mu}_\infty(\boldsymbol{\theta}), \boldsymbol{\Sigma}_\infty(\boldsymbol{\theta}))$, under the condition that

$$\boldsymbol{\mu}((\mathbf{X}_{n,j}^k)_{k\in[m]}) \underset{m\to\infty}{\longrightarrow} \boldsymbol{\mu}_\infty(\boldsymbol{\theta}_j), \text{ and}$$

$$\boldsymbol{\Sigma}((\mathbf{X}_{n,j}^k)_{k\in[m]}) \underset{m\to\infty}{\longrightarrow} \boldsymbol{\Sigma}_\infty(\boldsymbol{\theta}_j),$$

in measure $f_n(\cdot|\boldsymbol{\theta}_j)\mathrm{d}\lambda$, along with uniform integrability assumptions on the sequences of measures $(g_{n,m}(\cdot|\mathbf{x}_n)\mathrm{d}\lambda)_{m\in\mathbb{N}}$ and densities

$$(\mathcal{N}_d(\boldsymbol{\eta}(\mathbf{x}_n); \boldsymbol{\mu}((\mathbf{X}_{n,j}^k)_{k\in[m]}), \boldsymbol{\Sigma}((\mathbf{X}_{n,j}^k)_{k\in[m]})))_{m\in\mathbb{N}},$$

for each $j \in [N]$. We note that these results only say that $g_m(\cdot|\mathbf{x}_n)$ converges to some $g(\cdot|\mathbf{x}_n)$, as $m \to \infty$, but provide no intuition regarding the form of $g(\cdot|\mathbf{x}_n)$, nor how it relates to the target: $\pi(\cdot|\mathbf{x}_n)$.

Stronger results were obtained by [8], who required stricter conditions, to prove Bernstein–von Mises-type normal limit theorems for the class of covariance estimator-adjusted BSL techniques. For instance, the authors assume a central limit theorem with respect to the summary $\boldsymbol{\eta}(\mathbf{X}_n)$ and its limit $\boldsymbol{\eta}_0$, for some $\boldsymbol{\theta}_0 \in \mathbb{R}^d$. They further assume that there is a mapping $\boldsymbol{\theta} \mapsto \bar{\boldsymbol{\eta}}(\boldsymbol{\theta})$, for which $\bar{\boldsymbol{\eta}}(\boldsymbol{\theta}_0) = \boldsymbol{\eta}_0$, uniquely for some $\boldsymbol{\theta}_0 \in \mathbb{T}$, and further that $\bar{\boldsymbol{\eta}}$ is differentiable with full-rank Jacobian in a neighbourhood of $\boldsymbol{\theta}_0$. The covariance matrix estimator is further assumed to satisfy a uniform law of large numbers, under appropriate scaling, and the moment generating function of the scaled difference between $\boldsymbol{\eta}(\mathbf{X}_n)$ and $\boldsymbol{\eta}_0$, which admits a central limit theorem, also has sub-Gaussian tails for sufficiently large n. Under these assumptions, the BSL posterior density estimator obtained using covariance estimators that satisfy the regularity conditions will converge, in probability, to a normal density function, in the total variation topology, as both n and m approach infinity. This can better be interpreted as the convergence in distribution of an appropriately scaling of the posterior mean, $\mathbf{X}_n \mapsto \int_{\mathbb{T}} \boldsymbol{\theta} g_m(\boldsymbol{\theta}|\mathbf{X}_n)\lambda(\mathrm{d}\boldsymbol{\theta})$, to a normal random variable.

Notice that these assumptions are difficult to intuit and to verify for many sufficiently complex practical scenarios, and can be violated in simple cases. For example, one cannot use summary statistics such as M-estimator solutions [27], where the extrema are unidentifiable (see, *e.g.*, [26, Sec. 5.1]); nor U- and V-statistics defined by degenerate kernels [12, Ch. 4].

III. BSL VIA MIXTURE OF EXPERTS

A. Surrogate likelihoods via mixture of experts

As with the BSL methods described above, we seek to approximate the likelihood $f_n(\mathbf{x}_n|\boldsymbol{\theta})$ in some form, when $\mathbb{T} \subset \mathbb{R}^p$, for some $p \in \mathbb{N}$. Namely, given a choice of summary statistic $\boldsymbol{\eta} : \mathbb{X}^n \to \mathbb{R}^d$, we consider the classes of MoEs with normal experts and Gaussian gating (cf. [11], [15], [32]). The main reason for this choice is that the maximum conditional likelihood estimator (MCLE) is well approximated by the computationally more convenient GLLiM model estimator of [4]. *Mutatis mutandis*, the same results can be obtained for the softmax gating function via the equivalence between two classes (cf. [17, Lem. 1]). Writing $\boldsymbol{\eta}_n = \boldsymbol{\eta}(\mathbf{x}_n)$, our likelihood approximators take the form $(\boldsymbol{\eta}_n, \boldsymbol{\theta}) \mapsto h_{n,K}(\boldsymbol{\eta}_n|\boldsymbol{\theta}) \in \mathscr{M}_K$, where $K \in \mathbb{N}$ is the number of mixture components, and

$$\mathscr{M}_K = \left\{ h_K : \mathbb{R}^d \times \mathbb{T} \to \mathbb{R} : h_K(\boldsymbol{\eta}|\boldsymbol{\theta}) = \sum_{k=1}^{K} \gamma_k(\boldsymbol{\theta}; \psi_K) \mathcal{N}_d(\boldsymbol{\eta}; \boldsymbol{b}_k + \boldsymbol{A}_k\boldsymbol{\theta}, \boldsymbol{\Sigma}_k) \right\} \quad (4)$$

with $\boldsymbol{b}_k \in \mathbb{R}^d$, $\boldsymbol{A}_k \in \mathbb{R}^{d\times p}$, and $\boldsymbol{\Sigma}_k \in \mathcal{S}_d^+$ (the positive definite matrices in $\mathbb{R}^{d\times d}$), for each $k \in [K]$. Further, we take the sequence of gating functions $\boldsymbol{\gamma}_K(\cdot; \psi_K) = (\gamma_k(\cdot; \psi_K))_{k\in[K]}$ in the set of Gaussian gating functions:

$$\mathscr{G}_K = \left\{ \boldsymbol{\gamma}_K(\cdot; \psi_K) : \gamma_k(\boldsymbol{\theta}; \psi_K) = \frac{\pi_k \mathcal{N}_p(\boldsymbol{\theta}; \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)}{\sum_{l=1}^{K} \pi_l \mathcal{N}_p(\boldsymbol{\theta}; \boldsymbol{c}_l, \boldsymbol{\Gamma}_l)},\ k \in [K] \right\},$$

where $\boldsymbol{c}_k \in \mathbb{R}^p$, $\boldsymbol{\Gamma}_k \in \mathcal{S}_p^+$, and $\pi_k \in [0, 1]$, for each $k \in [K]$, with $\sum_{k=1}^K \pi_k = 1$. We denote $\psi_K = (\pi_k, \boldsymbol{c}_k, \boldsymbol{\Gamma}_k)_{k\in[K]}$, and $\boldsymbol{\chi}_K = (\boldsymbol{b}_k, \boldsymbol{A}_k, \boldsymbol{\Sigma}_k)_{k\in[K]}$. Then we assume that $\boldsymbol{\Psi}_K = (\psi_K, \boldsymbol{\chi}_K) \in \mathcal{X}$, for some domain $\mathcal{X}$ satisfying the parameter space restrictions, above.

For a fixed $\boldsymbol{\Psi}_K$, a mixture in $\mathcal{M}_K$ can be seen as a function of $\boldsymbol{\theta}$. The idea is then to learn an estimate of $\boldsymbol{\Psi}_K$ so that the corresponding mixture is a good approximation of the likelihood for every $\boldsymbol{\theta}$.

B. Approximation capacities

The approximation of the likelihood by a function in $\mathscr{M}_K$ is appealing for a number of reasons. Let $\bar{f}_n(\boldsymbol{\eta}_n|\boldsymbol{\theta})$ denote the pushforward likelihood of $\boldsymbol{\eta}_n = \boldsymbol{\eta}(\mathbf{X}_n)$, based on $f_n(\mathbf{x}_n|\boldsymbol{\theta})$. Then, on every compact subset $\mathbb{K} \subset \mathbb{T}$, as long as $\bar{f}_n(\boldsymbol{\eta}_n; \cdot)$ is continuous on $\mathbb{T}$, for every $\epsilon > 0$, there exists a sufficiently large $K \in \mathbb{N}$ and $h_{n,K}(\boldsymbol{\eta}_n|\boldsymbol{\theta}) \in \mathscr{M}_K$, such that the conditional expectations according to $\bar{f}_n(\boldsymbol{\eta}_n|\boldsymbol{\theta})$ and $h_{n,K}(\boldsymbol{\eta}_n|\boldsymbol{\theta})$ are uniformly close [16]:

$$\sup_{\boldsymbol{\theta}\in\mathbb{K}} \left| \int_{\mathbb{R}^d} \boldsymbol{\eta}_n \left\{ \bar{f}_n(\boldsymbol{\eta}_n|\boldsymbol{\theta}) - h_{n,K}(\boldsymbol{\eta}_n|\boldsymbol{\theta}) \right\} \lambda(\mathrm{d}\boldsymbol{\eta}_n) \right| < \epsilon.$$

This implies that we can approximate the mean of any pushforward likelihood arbitrarily well using an approximation in $\mathscr{M}_K$, for sufficiently large K. Further, on any compact sets $\mathbb{H} \subset \mathbb{R}^d$ and $\mathbb{K} \subset \mathbb{T}$, if $\bar{f}_n$ is a density on $\mathbb{H}$ for each fixed $\boldsymbol{\theta} \in \mathbb{K}$, and if $\bar{f}_n$ is continuous on $\mathbb{H} \times \mathbb{K}$, then, by [17, Thm. 1], for each $q \in [1, \infty)$ and $\epsilon > 0$, there exists a $h_{n,K}(\boldsymbol{\eta}_n|\boldsymbol{\theta}) \in \mathscr{M}_K$ such that

$$\left\{ \int_{\mathbb{H}\times\mathbb{K}} \left| \bar{f}_n(\boldsymbol{\eta}_n|\boldsymbol{\theta}) - h_{n,K}(\boldsymbol{\eta}_n|\boldsymbol{\theta}) \right|^q \lambda(\mathrm{d}\boldsymbol{\eta}_n \times \mathrm{d}\boldsymbol{\theta}) \right\}^{1/q} < \epsilon.$$

Thus, not only is the mean of $h_{n,K}(\boldsymbol{\eta}_n|\boldsymbol{\theta})$ close to its target, but if the target is compactly supported, then $h_{n,K}(\boldsymbol{\eta}_n|\boldsymbol{\theta})$ will be close to its target conditional density in any q-norm as well.

C. Posterior consistency

In Bayesian settings, the subsequent step is to consider the posterior distribution induced by the likelihood approximation. For fixed, $K \in \mathbb{N}$, we first wish to estimate the parameter $\boldsymbol{\Psi}_K$ that determines the *optimal* $h_{n,K} = h_{n,K}(\cdot|\cdot; \boldsymbol{\Psi}_K)$, where we now make the dependence on $\boldsymbol{\Psi}_K$ explicit. More specifically, using N simulated *i.i.d.* samples $\mathbf{Y}_N = ((\boldsymbol{\theta}_j, \mathbf{X}_{n,j}))_{j\in[N]}$ from the joint measure $\mathbb{Q}_n$ capturing the likelihood information, we consider the MCLE of $\boldsymbol{\Psi}_K$:

$$\hat{\boldsymbol{\Psi}}_{K,N} = \arg\max_{\boldsymbol{\Psi}_K} \frac{1}{N} \sum_{j=1}^{N} \log h_{n,K}(\boldsymbol{\eta}(\mathbf{X}_{n,j})|\boldsymbol{\theta}_j; \boldsymbol{\Psi}_K), \quad (5)$$

and parameters $\boldsymbol{\Psi}_K^*$ minimizing the Kullback–Leibler divergence between $\bar{f}_n$ and mixtures in $\mathcal{M}_K$ (cf. [30, Ch. 21]):

$$\min_{\boldsymbol{\Psi}_K} \mathbb{E}_{\mathbb{Q}_n} \log \left\{ \frac{\bar{f}_n\left(\boldsymbol{\eta}\left(\mathbf{X}_n\right)|\boldsymbol{\theta}\right)}{h_{n,K}\left(\boldsymbol{\eta}\left(\mathbf{X}_n\right)|\boldsymbol{\theta}; \boldsymbol{\Psi}_K\right)} \right\} = \mathbb{E}_{\mathbb{Q}_n} \log \left\{ \frac{\bar{f}_n(\boldsymbol{\eta}(\mathbf{X}_n)|\boldsymbol{\theta})}{h_{n,K}(\boldsymbol{\eta}(\mathbf{X}_n)|\boldsymbol{\theta}; \boldsymbol{\Psi}_K^*)} \right\}.$$

For each parameter $\boldsymbol{\Psi}_K$, we then define the posterior measure corresponding to $h_{n,K}(\boldsymbol{\eta}(\mathbf{x}_n)|\boldsymbol{\theta}; \boldsymbol{\Psi}_K)$ via the density

$$h_{n,K}(\boldsymbol{\theta}|\mathbf{x}_n; \boldsymbol{\Psi}_K) \propto h_{n,K}(\boldsymbol{\eta}(\mathbf{x}_n)|\boldsymbol{\theta}; \boldsymbol{\Psi}_K)\pi(\boldsymbol{\theta}).$$

and show the following convergence result.

Theorem 1 (Posterior consistency)**.** *Assume that $\boldsymbol{\eta}(\mathbf{X}_n)$ and $\boldsymbol{\theta}$ have finite second moments with respect to $\mathbb{Q}_n$ and $\mathcal{X}$ is compact. Then, if Φ is compact and $\left(\hat{\boldsymbol{\Psi}}_{K,N}\right)_{N\in\mathbb{N}}$ is a convergent sequence, the posterior measures defined by $\left(\hat{\boldsymbol{\Psi}}_{K,N}\right)_{N\in\mathbb{N}}$ converge in total variation, almost surely, to the posterior measure defined by $\boldsymbol{\Psi}_K^*$, in the sense that, for each $\mathbf{x}_n \in \mathbb{X}^n$,*

$$\int_{\mathbb{T}} \left| h_{n,K}\left(\boldsymbol{\theta}|\mathbf{x}_n; \hat{\boldsymbol{\Psi}}_{K,N}\right) - h_{n,K}\left(\boldsymbol{\theta}|\mathbf{x}_n; \boldsymbol{\Psi}_K^*\right) \right| \lambda\left(d\boldsymbol{\theta}\right) \underset{N\to\infty}{\longrightarrow} 0$$

for almost every $(\boldsymbol{Y}_N)_{N\in\mathbb{N}}$.

The proof of Theorem 1 is given in the Appendix.

D. Fast convergence rates

Note that not only is the MoE approximation of the likelihood and posterior consistency attractive, but we can also obtain near-optimal convergence estimation rates via Theorem 2, which is proved in the Appendix. We specialize to the well-specified case, where the generative measure $\mathbb{Q}_n$ has conditional density in $\mathscr{M}_K$, denoted as $h_{n,K^0}(\cdot|\cdot;\mathbf{\Psi}^0_{K^0})$ with K^0 number of mixture components, where $K_0 \leq K$. For each $\boldsymbol{\theta} \in \mathbb{T}$, we define the Hellinger distance, denoted by $\mathrm{He}(\cdot,\cdot)$, as follows:

$$\mathrm{He}\left(h_{n,K}(\cdot|\boldsymbol{\theta};\hat{\mathbf{\Psi}}_{K,N}), h_{n,K^0}(\cdot|\boldsymbol{\theta};\mathbf{\Psi}^0_{K^0})\right) = \left[\frac{1}{2}\int_{\mathcal{X}}\left(\sqrt{h_{n,K}(\boldsymbol{\eta}_n|\boldsymbol{\theta};\hat{\mathbf{\Psi}}_{K,N})} - \sqrt{h_{n,K^0}(\boldsymbol{\eta}_n|\boldsymbol{\theta};\mathbf{\Psi}^0_{K^0})}\right)^2 \times \lambda(\mathrm{d}\boldsymbol{\eta}_n)\right]^{1/2}.$$

Theorem 2 (Conditional density estimation)**.** *Assume that $((\boldsymbol{\theta}_j, \mathbf{X}_{n,j}))_{j\in[N]}$ are sampled i.i.d from generative joint measure $\mathbb{Q}_n$. Assume that $\mathcal{X}$ is compact and $\mathbb{T}$ is bounded. Given $\hat{\mathbf{\Psi}}_{K,N}$ defined in* (5)*, the corresponding conditional density function $h_{n,K}(\cdot|\cdot;\hat{\mathbf{\Psi}}_{K,N})$ admits the convergence rate of order $\mathrm{O}((\log N/N)^{-1/2})$ under the Hellinger distance in the sense that:*

$$\mathbb{P}\Big(\mathbb{E}_\Pi\left[\mathrm{He}\left(h_{n,K}(\cdot|\boldsymbol{\theta};\hat{\mathbf{\Psi}}_{K,N}), h_{n,K^0}(\cdot|\boldsymbol{\theta};\mathbf{\Psi}^0_{K^0})\right)\right] > C_1(\log N/N)^{-1/2}\Big) \leq C_2 N^{-C_3},$$

where C_1, C_2 and C_3 are universal positive constants.

IV. Numerical illustrations

A. Surrogate MoE likelihoods via GLLiM

For our numerical illustration, we use the GLLiM estimator of [4]. GLLiM has been used previously in [6] to provide surrogate posterior estimators. In our current setting, it is the likelihood that we approximate as an MoE:

$$h_{n,K}(\boldsymbol{\eta}_n|\boldsymbol{\theta};\mathbf{\Psi}_K) = \sum_{k=1}^{K}\gamma_k(\boldsymbol{\theta};\boldsymbol{\psi}_K)\mathcal{N}_d(\boldsymbol{\eta}_n;\boldsymbol{b}_k + \boldsymbol{A}_k\boldsymbol{\theta}, \mathbf{\Sigma}_k) \quad (6)$$

with $n = 1$ and $\boldsymbol{\eta}_n(\mathbf{X}_n) = X$ in each of our examples.

In the pBIL framework and notation of [5], we thus have an auxiliary model $h_{n,K}$, which can be viewed as a mixture of K Gaussian densities with parameters

$$\Phi(\boldsymbol{\theta};\mathbf{\Psi}_K) = ((\gamma_k(\boldsymbol{\theta};\boldsymbol{\psi}_K), \boldsymbol{b}_k + \boldsymbol{A}_k\boldsymbol{\theta}, \mathbf{\Sigma}_k))_{k\in[K]}. \quad (7)$$

Specifically, $\Phi(\boldsymbol{\theta}, \mathbf{\Psi}_K)$ is now a parametric function of $\boldsymbol{\theta}$ that depends on $\mathbf{\Psi}_K$ and specifies the proportions, means and covariance matrices of the K components. To define the mapping Φ, we only need an estimate of $\mathbf{\Psi}_K$. The parameter $\mathbf{\Psi}_K$ can be estimated, from the sample $\mathbf{Y}_N$, using a GLLiM model estimator $\bar{\mathbf{\Psi}}_{K,N}$, computed via a standard Expectation–Maximization (EM) algorithm. Details of the estimation procedure appears in [4]. Once we have computed $\bar{\mathbf{\Psi}}_{K,N}$, no further simulations are required. That is, $\mathbf{\Psi}_K$ can be estimated using only the size N simulation: $\mathbf{Y}_N = ((\boldsymbol{\theta}_j, \mathbf{X}_{n,j}))_{j\in[N]}$, with N fixed and independent of the required number of MCMC iterations. In the sequel, we will referred to our approach, using GLLiM model estimated MoE for BSL, as GLLiM-BSL.

B. Posterior samples

To sample from the posterior measure, BSL procedures use an estimation of the likelihood, plugged into an MCMC algorithm. In the BSL package [2], the default MCMC scheme is a Random Walk Metropolis Hastings (RW MH) algorithm, as provided by the mcmc package [9]. The covariance matrix of the Gaussian proposal is set to $s\boldsymbol{I}$, where $s > 0$ is a scale parameter that has to be carefully chosen and $\boldsymbol{I}$ is the identity matrix. For GLLiM-BSL, we also test a Slice Sampler (SS) [14] and a Metropolis Hastings scheme, using the GLLiM approximation of the posterior as a proposal distribution (G MH). These two latter choices have the advantage of not requiring tuning. For all MCMC schemes, we perform 3×10^5 iterations, with a burnin of 2×10^5 and a 1-in-100 sample thinning, resulting in a sample of 1000 $\boldsymbol{\theta}$ values.

An MoE is learned on a sample $\mathbf{Y}_N$ of size $N = 10^5$, obtained by simulating parameters from the prior and underlying measure defined by f_n, using a GLLiM estimator. The Bayesian information criterion (BIC) is used to choose the number of mixture components K. Once estimated with the selected K, the MoE provides an approximation of the likelihood which is used together with one of the aforementioned MCMC schemes. For comparison, we also use the GLLiM model approximation of the posterior measure to directly generate a sample of size 1000, as per [6]. This does not require any MCMC scheme. These direct GLLiM-based samples are then compared with samples resulting from various BSL procedures from the BSL package: BSL, semiBSL, missBSLmean, missBSLvar and uBSL; see [2] for details. We limit to visual comparison as it is enough to illustrate the improvement obtained by our method. There exists quantitative criteria for comparing samples, such as distances between samples (*e.g.* Wasserstein, energy distances etc.), 2-sample tests, *etc.* [13]. However, they provide highly volatile and inconsistent rankings between methods that are inconsistent with visual diagnoses. The development of quality assessment tools in likelihood free settings is actually an open question. It is a promising direction for future research that falls outside the scope of this paper.

C. Two moons example

The two moons model corresponds to a simulator that, given some parameters $\boldsymbol{\theta} = (\theta_1, \theta_2) \in \mathbb{R}^2$, produces an observation $\boldsymbol{X} \in \mathbb{R}^2$ via the scheme: $\boldsymbol{X} = \boldsymbol{P} + \frac{1}{\sqrt{2}}(-|\theta_1 + \theta_2|, -\theta_1 + \theta_2)$, with $\boldsymbol{P} = [R\cos(U) + 0.22, R\sin(U)]$ and $U \sim \mathcal{U}(-\pi/2, \pi/2)$, $R \sim \mathcal{N}(0.1, 0.01^2)$, where $\mathcal{U}$ is the uniform distribution.

We adopt the same setting as in [10]. Variable P follow a single crescent-shaped distribution, which is subsequently shifted and rotated around the origin, depending on $\boldsymbol{\theta}$. The absolute value $|\theta_1 + \theta_2|$ gives rise to the second crescent

in the posterior. The prior is uniform over $[-1,1]^2$ and the observed data is set to $\boldsymbol{x} = (0,0)$. The likelihood cannot be expressed explicitly but Figure 1 (a) shows 1000 simulations for $\boldsymbol{\theta} = (-0.5, 0.75)$, which clearly exhibit a non-Gaussian shape. A sample obtained from the GLLiM approximation of the likelihood, with $K = 49$ Gaussian components, is shown in Figure 1 (b), for comparison. The approximation is quite good, with a few extra outliers visible on the right indicating that some of the components are located there, but with low weight. The true posterior measure is made of two moon-like parts, see *e.g.* [10] and Figure 2 (a).

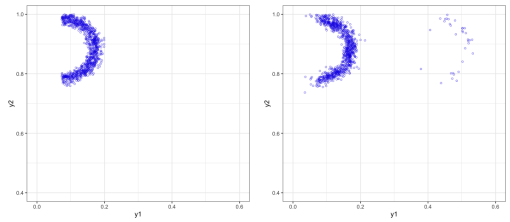


Fig. 1. Data $\boldsymbol{X}$ generated from the Two Moons example for $\boldsymbol{\theta} = (-0.5, 0.75)$. Samples of size 1000 from (a) the 2 moons simulator and from (b) the GLLiM likelihood estimation with $K = 49$ Gaussian components.

The GLLiM model is estimated using simulations $\mathbf{Y}_N$. Selecting from $K = 2$ to 50, the smallest BIC was obtained for $K = 49$. Figure 2 shows the different obtained samples. In this example, only the RW MH algorithm is tested as a MCMC scheme. We note that the uBSL variant from the BSL package exhibited a runtime error in this particular example and, as a result, was not used.

All methods identify the bimodality of the posterior distribution, but the BSL methods do not correctly recover the local structure of the two parts. In contrast, GLLiM-BSL provides a good representation of the posterior mass and moon structures. Among GLLiM-based procedures, the two moons were slightly better recovered with GLLiM-BSL than with the direct GLLiM posterior approximation. Table I shows the computing times obtained on a laptop (MacBook Pro, 2.4 GHz Quad-Core Intel Core i5) using the mentioned CRAN packages, and additional basic R code with no resorting to parallel computing. For the low dimensions of this example, *i.e.*, $d = p = 2$, the computing times were always less for GLLiM-based procedures, but not significantly so. However, the amortization nature of the GLLiM solution becomes an advantage in higher-dimensional problems, as seen in the following example.

D. Hyperboloids example

This example was introduced in [6] and exhibits a posterior distribution whose mass is located on 4 hyperboloids, as illustrated in Figure 3 (a). The GLLiM estimator was used to produce an MoE with $K = 38$ mixture components, as selected by BIC. The GLLiM-based likelihood was used with an RW MH algorithm to make comparisons with the standard BSL procedures. We also considered SS and G MH. In G MH, the variance of the GLLiM posterior was multiplied by 2 to avoid the proposal distribution being too narrow. The acceptance rate was 60% for G MH vs 16% for RW MH.

As depicted in Figure 3, although the posterior is far from being unimodal, some of the standard BSL variants (semiBSL and missBSL) succeed in capturing it satisfactorily compared to the previous example. This is likely due to the fact that the likelihood is simpler here, being a mixture of two Student distributions. Figure 3 shows the best results, obtained with GLLiM-BSL (c,d) and semiBSL (f). GLLiM approximations (Figure 3 (b,c,d)) appear to be better at capturing the hyperboloid branches, while some of the BSL variants (f,g,h), are more precise in the center with an obvious excess mass at the intersections of the branches. To complement this visual comparison, we also show the posterior marginals in Figure 4. The marginal plots allow us to better visualize the difference with standard BSL procedures. Refer to Figure 4 (f-j), which all show larger deviations from the truth, determined by numerical integration. Both true posterior marginals are the same due to symmetry in the model formulation and exhibit a non-smooth shape, which has also been double-checked using a long run of 3×10^5 iterations of the SS algorithm; see Figure 4 (a). For GLLiM-BSL, among the three MCMC schemes, it appears that the SS version in Figure 4 (d) provides more satisfactory samples than the MH versions (c, e). The gain over the direct GLLiM posterior sample (b) is also clearer. Computing times are reported in Table I. For the larger dimensional example ($d = 10$), GLLiM methods take much less time than standard BSL, even when considering BIC and learning times.

V. Conclusion

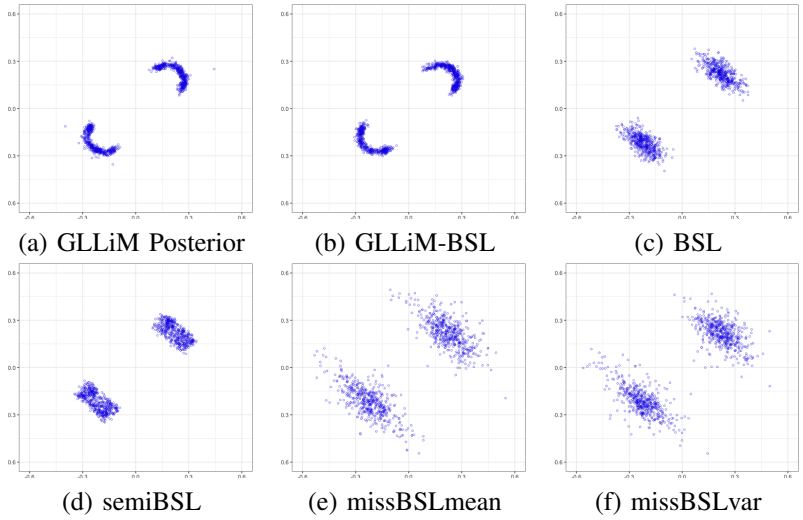
MoE approaches provide several advantages over previous BSL variants. The flexibility of the model allows for better approximations of likelihoods that strongly depart from Gaussianity. In particular, GLLiM model estimators have interesting amortization properties. GLLiM-based procedures can be applied in a wide variety of settings, such as sets of *i.i.d.* observations or time series, as illustrated in [6].To the best of our knowledge, we are the first to demonstrate approximation and estimation theoretical properties of MoEs in this setting.

VI. Appendix

Proof of Theorem 1. By the uniform strong law of large numbers [26, Thm. 9.60],

$$\sup_{\boldsymbol{\Psi}_K \in \mathcal{X}} \Big| N^{-1} \sum_{j=1}^{N} \log h_{n,K}\left(\boldsymbol{\eta}\left(\mathbf{X}_{n,j}\right)|\boldsymbol{\theta}_j; \boldsymbol{\Psi}_K\right) - \mathbb{E}_{\mathbb{Q}_n}\left[\log h_{n,K}\left(\boldsymbol{\eta}\left(\mathbf{X}_{n,j}\right)|\boldsymbol{\theta}_j; \boldsymbol{\Psi}_K\right)\right] \Big| \xrightarrow[N\to\infty]{\text{a.s.}} 0$$

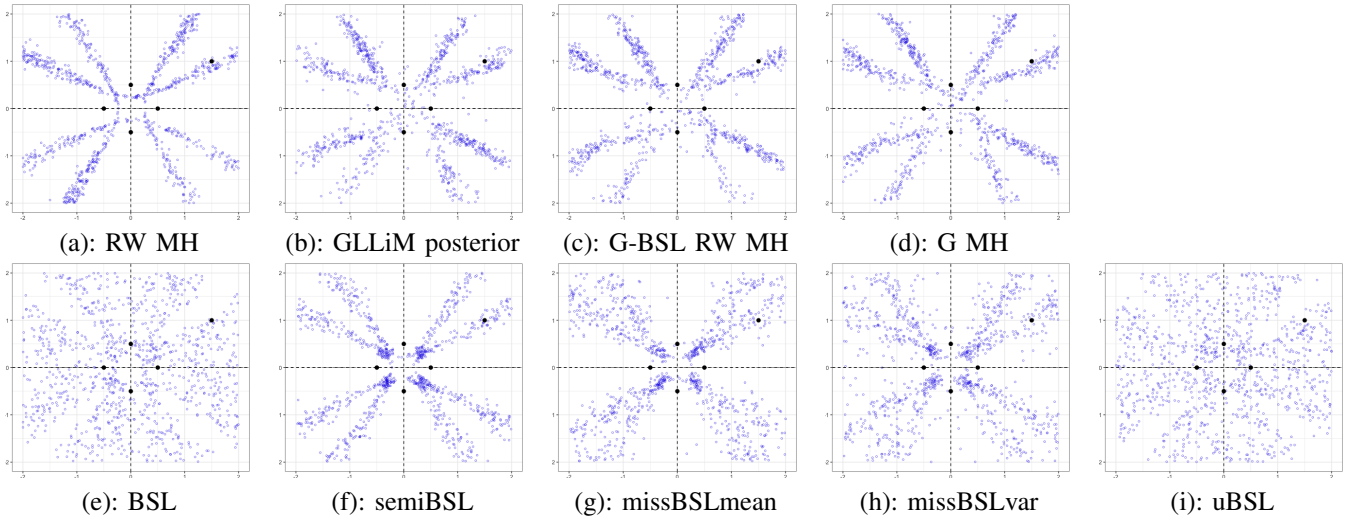
under the assumptions that $\boldsymbol{\eta}(\mathbf{X}_n)$ and $\boldsymbol{\theta}$ have finite second moments with respect to $\mathbb{Q}_n$, and compact $\mathcal{X}$. By [26, Thm. 5.3], this implies that since $\left(\hat{\boldsymbol{\Psi}}_{K,N}\right)_{N\in\mathbb{N}}$ is convergent and conditional likelihood maximizing, $\hat{\boldsymbol{\Psi}}_{K,N} \xrightarrow[N\to\infty]{} \boldsymbol{\Psi}_K^*$ for almost every $(\boldsymbol{Y}_N)_{N\in\mathbb{N}}$, for some Kullback–Leibler divergence



(a) GLLiM Posterior (b) GLLiM-BSL (c) BSL

(d) semiBSL (e) missBSLmean (f) missBSLvar

Fig. 2. Two moons example. Plots are zoomed in on $[-0.6, 0.6]^2$. Plots (a) and (b): GLLiM posterior and GLLiM-BSL samples for $K = 49$. Plots (c) to (f): BSL variants, respectively BSL, semiBSL, missBSLmean and missBSLvar. The MCMC scheme is a random walk Metropolis Hastings algorithm.



(a): RW MH (b): GLLiM posterior (c): G-BSL RW MH (d): G MH

(e): BSL (f): semiBSL (g): missBSLmean (h): missBSLvar (i): uBSL

Fig. 3. Hyperboloid example. Plot (a): reference Metropolis Hastings (RW MH) sample. Plots (b,c,d): GLLiM posterior, GLLiM-BSL (RW MH and G MH) samples for $K = 38$. Plots (e) to (i): BSL variants with RW MH, respectively BSL, semiBSL, missBSLmean, missBSLvar, uBSL.

minimizing $\boldsymbol{\Psi}_K^*$. Then, on this almost sure event, the continuity of $h_{n,K}\left(\boldsymbol{\eta}\left(\mathbf{x}_n\right)|\boldsymbol{\theta};\cdot\right)$ implies that

$$h_{n,K}\left(\boldsymbol{\eta}\left(\mathbf{x}_n\right)|\boldsymbol{\theta};\hat{\boldsymbol{\Psi}}_{K,N}\right) \longrightarrow h_{n,K}\left(\boldsymbol{\eta}\left(\mathbf{x}_n\right)|\boldsymbol{\theta};\boldsymbol{\Psi}_K^*\right)$$

for every fixed $\mathbf{x}_n$ and $\boldsymbol{\theta}$. It suffices to show that on the event,

$$\int_{\mathbb{T}} h_{n,K}\left(\boldsymbol{\eta}\left(\mathbf{x}_n\right)|\boldsymbol{\theta};\hat{\boldsymbol{\Psi}}_{K,N}\right)\pi\left(\boldsymbol{\theta}\right)\lambda\left(\mathrm{d}\boldsymbol{\theta}\right) \underset{N\to\infty}{\longrightarrow} \int_{\mathbb{T}} h_{n,K}\left(\boldsymbol{\eta}\left(\mathbf{x}_n\right)|\boldsymbol{\theta};\boldsymbol{\Psi}_K^*\right)\pi\left(\boldsymbol{\theta}\right)\lambda\left(\mathrm{d}\boldsymbol{\theta}\right),$$

which follows from the dominated convergence theorem by noticing that, for each fixed $\mathbf{x}_n$ and $\boldsymbol{\Psi}_K$,

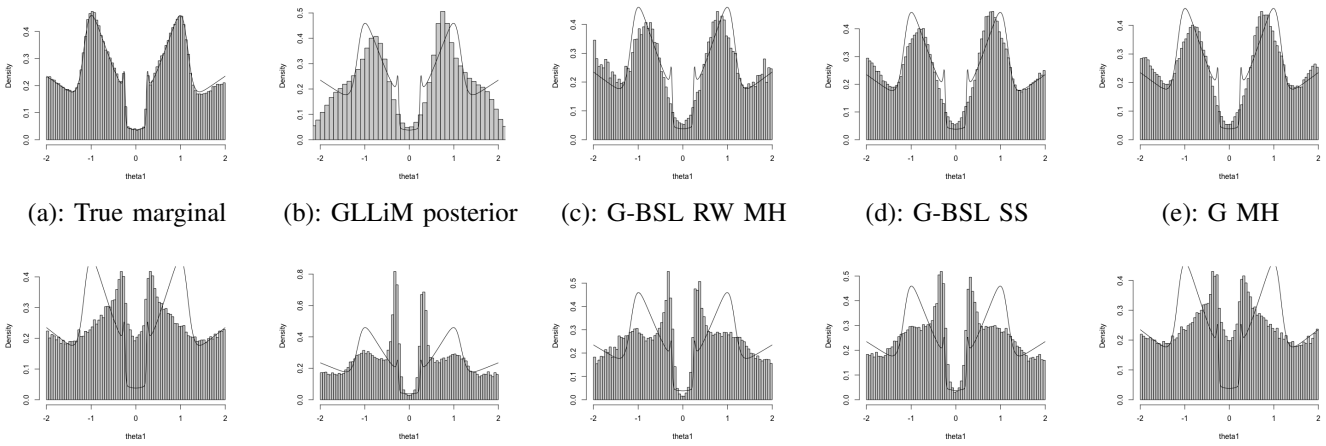
$$\left|h_{n,K}\left(\boldsymbol{\eta}\left(\mathbf{x}_n\right)|\boldsymbol{\theta};\boldsymbol{\Psi}_K\right)\pi\left(\boldsymbol{\theta}\right)\right| \leq C\pi\left(\boldsymbol{\theta}\right)$$

for some constant $C < \infty$ and $\int_{\mathbb{T}}\pi\left(\boldsymbol{\theta}\right)\lambda\left(\mathrm{d}\boldsymbol{\theta}\right) = 1$. We obtain our desired conclusion by an application of Scheffe's theorem [28, Cor. 2.30]. □

Proof of Theorem 2. It is convenient to index the true conditional density $h_{n,K^0}(\cdot|\cdot;\boldsymbol{\Psi}_{K^0}^0)$ as $h_{G^0}(\cdot|\cdot;\boldsymbol{\Psi}_{K^0}^0)$, by the discrete mixing measure on the parameters as follows: $G^0 = \sum_{k=1}^{K^0}\pi_k^0\delta_{(\boldsymbol{c}_k^0,\boldsymbol{\Gamma}_k^0,\boldsymbol{A}_k^0,\boldsymbol{b}_k^0,\boldsymbol{\Sigma}_k^0)}$ where $\delta_{(\boldsymbol{c}_k^0,\boldsymbol{\Gamma}_k^0,\boldsymbol{A}_k^0,\boldsymbol{b}_k^0,\boldsymbol{\Sigma}_k^0)}$ is the Dirac measure indexing the atom $(\boldsymbol{c}_k^0,\boldsymbol{\Gamma}_k^0,\boldsymbol{A}_k^0,\boldsymbol{b}_k^0,\boldsymbol{\Sigma}_k^0)$, for each $k \in [K^0]$. Here we denote the space of measures with at least K^0 atoms by $\mathcal{O}_K$, which equals

$$\left\{G = \sum_{k=1}^{\overline{K}}\pi_k\delta_{(\boldsymbol{c}_k,\boldsymbol{\Gamma}_k,\boldsymbol{A}_k,\boldsymbol{b}_k,\boldsymbol{\Sigma}_k)} : \overline{K} \in [K],\ K \geq K^0\right\}.$$

Note that for any $K^0 \leq K$, $\mathcal{O}_K$ can be defined equivalently as $\mathscr{M}_K = \{h_G(\boldsymbol{\eta}|\boldsymbol{\theta}) : G \in \mathcal{O}_K\}$ and write $\mathcal{Q}_K^{1/2} = \{h_{(G+G_0)/2}^{1/2}(\boldsymbol{\eta}|\boldsymbol{\theta}) : G \in \mathcal{O}_K\}$. Then, we define the Hellinger ball centered around the conditional density $h_{G^0}(\boldsymbol{\eta}|\boldsymbol{\theta})$ and intersected with the set $\mathcal{Q}_K^{1/2}$ by $\mathcal{Q}_K^{1/2}(\gamma) = \left\{g^{1/2} \in \mathcal{Q}_K^{1/2} :\right.$



(a): True marginal (b): GLLiM posterior (c): G-BSL RW MH (d): G-BSL SS (e): G MH

(f): BSL (g): semiBSL (h) : missBSLmean (i): missBSLvar (j): uBSL

Fig. 4. Hyperboloid posterior marginals. Plot (a): true marginal and slice sampler (SS) histogram. Plot (b): GLLiM posterior. Plots (c,d,e): GLLiM-BSL resp. with RW MH, SS and GLLiM posterior proposal (G MH). Plots (f) to (j): BSL variants, respectively BSL, semiBSL, missBSLmean, missBSLvar, uBSL.

TABLE I
SETTINGS AND COMPUTATION TIMES FOR THE 2 EXAMPLES AND VARIOUS PROCEDURES.

Example	Procedure	MCMC	p	d	K	N	m	BIC	GLLiM	$3 * 10^5$ iterations	R Package(s)
2 Moons	GLLiM BSL	RW MH	2	2	49	10^5	-	1h 28min	3min 6s	12min 30s	xLLiM, mcmc
	GLLiM post	-	2	2	49	10^5	-	1h 28min	3min 6s	-	xLLiM
	BSL	RW MH	2	2	-	-	500	-	-	23min 39s	BSL
	semiBSL	RW MH	2	2	-	-	500	-	-	33min 40s	BSL
	missBSLmean	RW MH	2	2	-	-	500	-	-	30min 21s	BSL
	missBSLvar	RW MH	2	2	-	-	500	-	-	29min 14s	BSL
Hyperboloids	GLLiM BSL	RW MH	2	10	38	10^5	-	1h 43min	4min 47s	43min 20s	xLLiM, mcmc
	GLLiM BSL	SS	2	10	38	10^5	-	1h 43min	4min 47s	2h 35min	xLLiM, diversitree
	GLLiM BSL	G MH	2	10	38	10^5	-	1h 43min	4min 47s	46min 28s	xLLiM
	GLLiM post	-	2	10	38	10^5	-	1h 43min	4min 47s	-	xLLiM
	BSL	RW MH	2	10	-	-	500	-	-	4h 19min	BSL, mcmc
	semiBSL	RW MH	2	10	-	-	500	-	-	4h 49min	BSL, mcmc
	missBSLmean	RW MH	2	10	-	-	500	-	-	4h 49min	BSL, mcmc
	missBSLvar	RW MH	2	10	-	-	500	-	-	4h 34min	BSL, mcmc
	uBSL	RW MH	2	10	-	-	500	-	-	4h 10min	BSL, mcmc

N is the number of samples used to learn a MoE and m is the number of simulations at each BSL iteration. The BIC column indicates the learning time for all GLLiM models between $K = 2$ and some $K_{\max}$, while the GLLiM column shows the time for the selected K indicated under column K. The second last column shows times for 3×10^5 MCMC iterations. The CRAN packages used are indicated in the last column.

$\mathrm{He}(g, h_{G^0}) \leq \gamma\Big\}$. Following the framework from [27], we introduce the following quantity to capture the size of the above Hellinger ball:

$$\mathcal{J}_B(\gamma, \mathcal{Q}_K^{1/2}) = \left[\int_{\gamma^2/2^{13}}^{\gamma} H_B^{1/2}(u, \mathcal{Q}_K^{1/2}(u), \|\cdot\|)\lambda(du)\right] \vee \gamma. \quad (8)$$

Here, $H_B^{1/2}(u, \mathcal{Q}_K^{1/2}(u), \|\cdot\|)$ denotes the bracketing entropy of $\mathcal{Q}_K^{1/2}(u)$ under the Euclidean distance, and $u \vee \gamma = \max\{u, \gamma\}$. Next, we introduce the upper bounds of the covering number (under the sup norm $\|\cdot\|_\infty$), $N(\epsilon, \mathscr{M}_K, \|\cdot\|_\infty)$, and the bracketing entropy (under the Hellinger distance) $H_B(\epsilon, \mathscr{M}_K, \mathrm{He})$ of the metric space $\mathscr{M}_K$. Note that by using the definition of the spaces $\mathcal{Q}_K^{1/2}$ and $\mathscr{M}_K$ and the relationship between $\|\cdot\|$ and He, for any $u > 0$, it holds that

$$H_B^{1/2}(u, \mathcal{Q}_K^{1/2}(u), \|\cdot\|) \leq H_B^{1/2}(u, \mathscr{M}_K, \mathrm{He}). \quad (9)$$

Then (8) implies that $\mathcal{J}_B(\gamma, \mathcal{Q}_K^{1/2})$ is upper bounbed by

$$\int_{\gamma^2/2^{13}}^{\gamma} H_B^{1/2}(u, \mathscr{M}_K, \mathrm{He})\lambda(du) \vee \gamma \leq \int_{\gamma^2/2^{13}}^{\gamma} \log(1/u)\lambda(du) \vee \gamma \leq T(\gamma). \quad (10)$$

The first inequality follows from Lemmas 4.3 and 4.6 in [21] while the second inequality is obtained with $T(\gamma) = \gamma[\log(1/\gamma)]^{1/2}$ and that $T(\gamma)/\gamma^2$ is a non-increasing function of γ. Finally, let $\gamma_N = \sqrt{\log(N)/N}$, then $\sqrt{N}\gamma_N^2 \geq CT(\gamma_N)$ holds for some universal constant C. This leads to the desired convergence rate thanks to Lemma 1. The proof of Lemma 1 for the conditional density estimation rate is similar to Theorem 7.4 in [27] for joint densities and is not presented here.

Lemma 1 (Theorem 7.4 in [27])**.** *Take $T(\gamma) \geq \mathcal{J}_B(\gamma, \mathcal{Q}_K^{1/2})$ such that $T(\gamma)/\gamma^2$ is a non-increasing function of γ. Then,*

for a universal constant C and a sequence (γ_N) that satisfies $\sqrt{N}\gamma_N^2 \geq CT(\gamma_N)$, for $\gamma \geq \gamma_N$, it holds

$$\mathbb{P}\Big(\mathbb{E}_\Pi\left[\mathrm{He}\left(h_{n,K}(\cdot|\boldsymbol{\theta};\hat{\boldsymbol{\Psi}}_{K,N}), h_{n,K^0}(\cdot|\boldsymbol{\theta};\boldsymbol{\Psi}^0_{K^0})\right)\right] > \gamma\Big) \leq C\exp\left(-\frac{N\gamma^2}{C^2}\right).$$

□

REFERENCES

[1] Z. An, D. J. Nott, and C. Drovandi, "Robust Bayesian synthetic likelihood via a semi-parametric approach," *Statistics and Computing*, vol. 30, pp. 543–557, 2020.

[2] Z. An, L. F. South, and C. Drovandi, "BSL: An R Package for Efficient Parameter Estimation for Simulation-Based Models via Bayesian Synthetic Likelihood," 2019.

[3] Z. An, L. F. South, D. J. Nott, and C. C. Drovandi, "Accelerating Bayesian Synthetic Likelihood With the Graphical Lasso," *Journal of Computational and Graphical Statistics*, pp. 471–475, 2019.

[4] A. Deleforge, F. Forbes, and R. Horaud, "High-dimensional regression with Gaussian mixtures and partially-latent response variables," *Statistics and Computing*, vol. 25, pp. 893–911, Sep. 2015.

[5] C. Drovandi, A. Pettitt, and A. Lee, "Bayesian indirect inference using a parametric auxiliary model," *Statistical Science*, vol. 30, pp. 72–95, 2014.

[6] F. Forbes, H. D. Nguyen, T. T. Nguyen, and J. Arbel, "Summary statistics and discrepancy measures for ABC via surrogate posteriors," *Statistics and Computing*, vol. 32, 2022.

[7] D. T. Frazier and C. Drovandi, "Robust approximate Bayesian inference with synthetic likelihood," *Journal of Computational and Graphical Statistics*, vol. 30, pp. 958–976, 2021.

[8] D. T. Frazier, D. J. Nott, C. Drovandi, and R. Kohn, "Bayesian inference using synthetic likelihood: asymptotics and adjustments," *Journal of the American Statistical Association*, pp. 1–12, 2022.

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

[10] D. Greenberg, M. Nonnenmacher, and J. Macke, "Automatic posterior transformation for likelihood-free inference," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 2404–2414.

[11] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, pp. 79–87, 1991, publisher: MIT Press.

[12] V. S. Korolyuk and Y. V. Borovskich, *Theory of U-statistics*. Springer, 2013, vol. 273.

[13] J.-M. Lueckmann, J. Boelts, D. Greenberg, P. Goncalves, and J. Macke, "Benchmarking simulation-based inference," in *Proceedings 24th International Conference on Artificial Intelligence and Statistics*, 2021.

[14] R. M. Neal, "Slice sampling," *The Annals of Statistics*, vol. 31, pp. 705–767, 2003.

[15] H. D. Nguyen and F. Chamroukhi, "Practical and theoretical aspects of mixture-of-experts modeling: An overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, 2018.

[16] H. D. Nguyen, F. Chamroukhi, and F. Forbes, "Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model," *Neurocomputing*, vol. 366, pp. 208–214, 2019.

[17] H. D. Nguyen, T. Nguyen, F. Chamroukhi, and G. J. McLachlan, "Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models," *Journal of Statistical Distributions and Applications*, vol. 8, pp. 1–15, 2021.

[18] H. Nguyen, P. Akbarian, T. Nguyen, and N. Ho, "A General Theory for Softmax Gating Multinomial Logistic Mixture of Experts," *arXiv preprint arXiv:2310.14188*, 2023.

[19] H. Nguyen, T. Nguyen, and N. Ho, "Demystifying Softmax Gating Function in Gaussian Mixture of Experts," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[20] H. Nguyen, T. Nguyen, K. Nguyen, and N. Ho, "Towards Convergence Rates for Parameter Estimation in Gaussian-gated Mixture of Experts," *arXiv preprint arXiv:2305.07572*, 2023.

[21] T. Nguyen, H. D. Nguyen, F. Chamroukhi, and F. Forbes, "A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models," *Electronic Journal of Statistics*, vol. 16, pp. 4742 – 4822, 2022.

[22] V. Ong, D. Nott, M.-N. Tran, S. Sisson, and C. Drovandi, "Likelihood-free inference in high dimensions with synthetic likelihood," *Computational Statistics and Data Analysis*, vol. 128, 07 2018.

[23] L. F. Price, C. C. Drovandi, A. Lee, and D. J. Nott, "Bayesian synthetic likelihood," *Journal of Computational and Graphical Statistics*, vol. 27, pp. 1–11, 2018.

[24] J. W. Priddle, S. A. Sisson, D. T. Frazier, I. Turner, and C. Drovandi, "Efficient Bayesian Synthetic Likelihood With Whitening Transformations," *Journal of Computational and Graphical Statistics*, vol. 31, pp. 50–63, Jan. 2022.

[25] C. P. Robert, *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer, 2007, vol. 2.

[26] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on stochastic programming: modeling and theory*. SIAM, 2021.

[27] S. van de Geer, *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

[28] A. W. Van der Vaart, *Asymptotic statistics*. Cambridge University Press, 2000.

[29] S. Watanabe, *Mathematical theory of Bayesian statistics*. CRC Press, 2018.

[30] H. White, *Estimation, inference and specification analysis*. Cambridge university press, 1996.

[31] S. N. Wood, "Statistical inference for noisy nonlinear ecological dynamic systems," *Nature*, vol. 466, pp. 1102–1104, 2010.

[32] L. Xu, M. Jordan, and G. E. Hinton, "An Alternative Model for Mixtures of Experts," in *Advances in Neural Information Processing Systems*, vol. 7. MIT Press, 1995.

[33] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE transactions on neural networks and learning systems*, vol. 23, pp. 1177–1193, 2012.