Electronic Journal of Statistics
Vol. 18 (2024) 3628–3686
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2283

Generalized Bayesian likelihood-free inference

Lorenzo Pacchiardi[1], **Sherman Khoo**[2] **and Ritabrata Dutta**[3]

[1] *Leverhulme Centre for the Future of Intelligence, University of Cambridge, UK,*
e-mail: lp666@cam.ac.uk

[2] *School of Mathematics, University of Bristol, UK,*
e-mail: fo23878@bristol.ac.uk

[3] *Department of Statistics, University of Warwick, UK,*
e-mail: ritabrata.dutta@warwick.ac.uk

Abstract: Generalized Bayesian inference replaces the likelihood in the Bayesian posterior with the exponential of a loss function connecting parameter values and observations. As a loss function, it is possible to use Scoring Rules (SRs), which evaluate the match between the observation and the probabilistic model for given parameter values. In this work, we leverage this *Scoring Rule posterior* for Bayesian Likelihood-Free Inference (LFI). In LFI, we can sample from the model but not evaluate the likelihood; hence, we use the Energy and Kernel SRs in the SR posterior, as they admit unbiased empirical estimates. While traditional Pseudo-Marginal (PM) Markov Chain Monte Carlo (MCMC) can be applied to the SR posterior, it mixes poorly for concentrated targets, such as those obtained with many observations. As such, we propose to use Stochastic Gradient (SG) MCMC, which improves performance over PM-MCMC and scales to higher-dimensional setups as it is rejection-free. SG-MCMC requires differentiating the simulator model; we achieve this effortlessly by implementing the simulator models using automatic differentiation libraries. We compare SG-MCMC sampling for the SR posterior with related LFI approaches and find that the former scales to larger sample sizes and works well on the raw data, while other methods require determining suitable summary statistics. On a chaotic dynamical system from meteorology, our method even allows inferring the parameters of a neural network used to parametrize a part of the update equations.

MSC2020 subject classifications: Primary 62-08; secondary 62A01.
Keywords and phrases: Likelihood-free inference, generalized Bayes, scoring rules, pseudo-marginal MCMC.

Received April 2023.

Contents

1. Introduction

This work is concerned with performing inference for a model P_θ whose density $p(y|\theta)$ for an observation y is unavailable, but from which it is easy to simulate for any parameter value θ (such models are known as intractable-likelihood or simulator models). Given y and a prior $\pi(\theta)$ on the parameters, the standard Bayesian posterior is $\pi(\theta|y) \propto \pi(\theta)p(y|\theta)$. However, obtaining that explicitly or sampling from it with Markov Chain Monte Carlo (MCMC) techniques is impossible without having access to the likelihood.

Traditional Likelihood-Free Inference (LFI) techniques exploit model simulations to approximate the exact posterior distribution when the likelihood is unavailable, by either estimating an explicit surrogate [71, 1, 77] or weighting different parameter values according to the mismatch between observed and simulated data [50, 6].

In this work, we introduce a new LFI approach grounded in the generalized Bayesian inference framework [10, 41, 45]: given a generic loss $\ell(y, \theta)$ between a single observation y and parameter θ, the generalized posterior belief on parameter values can be defined as:

$$\pi(\theta|y) \propto \pi(\theta) \exp(-w \cdot \ell(y, \theta)); \tag{1}$$

this allows us to learn about the parameter value minimizing the expected loss over the data generating process[1] and respects Bayesian additivity (namely, the final posterior distribution does not depend on the order observations are received). The learning rate w controls how much the posterior concentrates when increasing the number of observed samples n.

Previous works [34, 52] took $\ell(y, \theta)$ to be a Scoring Rule (SR) $S(P_\theta, y)$, which assesses the performance of P_θ for an observation y. Here, we apply this *scoring rule posterior* π_S to Bayesian LFI. In particular, we consider scoring rules S such that $S(P_\theta, y)$ can be estimated with samples from P_θ; thus, we can perform LFI without worrying about the missing likelihood $p(y|\theta)$. Two scoring rules allowing this while having good theoretical properties are the kernel and the energy scores [35]. When $k(\cdot, \cdot)$ is a symmetric and positive-definite kernel, the kernel score for k can be defined as [35]:

$$S_k(P, y) = \mathbb{E}[k(X, X')] - 2 \cdot \mathbb{E}[k(X, y)], \quad X \perp\!\!\!\perp X' \sim P.$$

The energy score is given by a specific choice of k [35]:

$$S_{\mathrm{E}}^{(\beta)}(P, y) = 2 \cdot \mathbb{E}\left[\|X - y\|_2^\beta\right] - \mathbb{E}\left[\|X - X'\|_2^\beta\right], \quad X \perp\!\!\!\perp X' \sim P,$$

where $\beta \in (0, 2)$.

[1] Indeed setting $\ell(y, \theta) = -\log p(y|\theta)$ and $w = 1$ recovers the standard Bayes update, which learns about the parameter value minimizing the KL divergence [10].

When inserting the kernel score in Eq. (1), the MMD-Bayes posterior [13] is recovered. The SR posterior can be therefore seen as a generalization of MMD-Bayes, which was also employed for LFI in the original work [13]. In the present paper, we discuss the SR posterior for LFI in more generality and employ MCMC schemes to perform inference (instead of variational inference as in [13]).

Exact sampling from the SR posterior remains impossible; still, a Pseudo-Marginal (PM) MCMC [4] where simulations from $P_{\theta'}$ are generated for each proposed θ' can be used to sample from a close approximation (whose error diminishes when the number of simulations at each step increases) for any SR allowing estimation from samples. While PM-MCMC works well for simple cases and is applicable to any simulator model, it mixes poorly for concentrated targets (such as those obtained when many observations are used).

Alternatively, approximate samples from the SR posterior can be obtained using Stochastic-Gradient (SG) MCMC [59] by leveraging the unbiased estimates of $\nabla_\theta S(P_\theta, y)$ possible with the energy and the kernel score. The unbiased gradient estimate requires the gradient of the simulated data with respect to model parameters, which can easily be obtained by implementing the simulator model with automatic differentiation libraries. In this work, we mostly employ adaptive stochastic gradient Langevin dynamics [43], which enjoys theoretical bounds for its error and results for asymptotic convergence [22, 47, 46]; further, we show empirically that the SG-MCMC target matches well that obtained with PM-MCMC in cases where the latter mixes well while requiring lower computational effort. Importantly, SG-MCMC has no mixing issues (as it is rejection-free). To the best of our knowledge, ours is the first application of gradient-based sampling methods to LFI using an unbiased estimate of the gradient of the target distribution, which is enabled by the SR posterior and leads to scalable inference for high-dimensional parameter spaces.

Qualitatively, the properties of the SR posterior are independent of the value of w in its definition (see Eq. 2). However, the choice of w determines the rate of contraction of the SR posterior. A large ongoing research effort is devoted to the selection of w for generalized Bayesian posteriors, resulting in methods ensuring, for instance, different forms of coverage [54, 76, 57] or other properties [10, 39, 52]. Several of those methods (and plausibly future ones) apply to our framework. Hence, we do not delve deep into determining the optimal way to select w or develop our own, mindful of the fact that this is an area of active research and that each practical use case is best tackled with a different method. Still, in our empirical evaluations of the SR posterior, it may be beneficial for different posteriors to have similar scales. When that is required, we will either rely on hand-tuning or a previously introduced method which we revisit for our framework.

We empirically compare the SR posterior with the popular Bayesian Synthetic Likelihood (BSL, 71) approach, which also involves estimating the posterior at each MCMC step via model simulations. However, as BSL does not provide unbiased gradient estimates, this prevents the use of SG-MCMC, which hinders the performance of BSL for concentrated and high-dimensional targets. Next, we consider a real-world meteorological model [53] and infer its param-

eters with Approximate Bayesian Computation [50] and our SR posterior. We also use our framework to infer the posterior distribution over the parameters of a high-dimensional Neural Stochastic-Differential Equation for modelling the same data, which is unachievable with traditional (non-gradient-based) sampling methods. Moreover, our method works on the raw data, whereas traditional LFI methods require determining suitable summary statistics.

To summarise, our contributions are as follows:

- We apply the Scoring Rule posterior [34, 52] to Likelihood-Free Inference (LFI), study its properties and discuss how it generalizes some existing LFI methods.
- We leverage stochastic gradient MCMC [59] for sampling from the Scoring Rule posterior by relying on automatic differentiation of the simulator models, and show how it performs better than pseudo-marginal MCMC.
- We conduct simulation studies where we compare existing LFI methods with our approach, which scales to higher-dimensional parameter space and a larger number of observations, mainly thanks to employing a gradient-based sampling method enabled by the SR posterior.

The rest of this manuscript is organized as follows. In Sec. 2, we first review the scoring rules and define the SR posterior; we then discuss and compare the two sampling methods in Section 3. Simulation studies assessing the performance of our proposed sampling scheme for scoring rule posterior and comparison with other LFI approaches are presented in Sec. 4. Finally, we briefly review previous works in Sec. 5 and conclude and suggest future directions in Sec. 6.

1.1. Notation

We will denote respectively by $\mathcal{X} \subseteq \mathbb{R}^d$ and $\Theta \subseteq \mathbb{R}^p$ the data and parameter space, which we assume to be Borel sets. We will assume the observations are generated by a distribution P_0 and use P_θ and $p(\cdot|\theta)$ to denote the distribution and likelihood of our model. Generic distributions will be indicated by P or Q, while S will denote a generic scoring rule. Other upper-case letters will denote random variables while lower-case ones will denote observed (fixed) values. We will denote by Y or y the observations (correspondingly random variables and realizations) and X or x the simulations. Subscripts will denote sample index and superscripts vector components. Also, we will respectively denote by $\mathbf{Y_n} = \{Y_i\}_{i=1}^n \in \mathcal{X}^n$ and $\mathbf{y_n} = \{y_i\}_{i=1}^n \in \mathcal{X}^n$ a set of random and fixed observations. Similarly, $\mathbf{X_m} = \{X_j\}_{j=1}^m \in \mathcal{X}^m$ and $\mathbf{x_m} = \{x_j\}_{j=1}^m \in \mathcal{X}^m$ denote a set of random and fixed model simulations. Finally, $\perp\!\!\!\perp$ will denote independence between random variables, while $X \sim P$ indicates a random variable distributed according to P.

2. Bayesian inference using scoring rules

2.1. Background definitions

A Scoring Rule (SR, 35) S is a function of a probability distribution over $\mathcal{X}$ and of an observation in $\mathcal{X}$. For a distribution P and an observation y, we will denote this as $S(P, y)$. Assuming that y is a realization of a random variable Y with distribution Q, the expected scoring rule is defined as:

$$S(P, Q) := \mathbb{E}_{Y \sim Q} S(P, Y),$$

where we overload notation in the second argument of S. The scoring rule S is *proper* relative to a set of distributions $\mathcal{P}(\mathcal{X})$ over $\mathcal{X}$ if

$$S(Q, Q) \leq S(P, Q) \; \forall \; P, Q \in \mathcal{P}(\mathcal{X}),$$

i.e., if the expected scoring rule is minimized in P when $P = Q$. Moreover, S is *strictly proper* relative to $\mathcal{P}(\mathcal{X})$ if $P = Q$ is the unique minimum:

$$S(Q, Q) < S(P, Q) \; \forall \; P, Q \in \mathcal{P}(\mathcal{X}) \text{ s.t. } P \neq Q.$$

The divergence related to a proper scoring rule [18] can be defined as $D(P, Q) := S(P, Q) - S(Q, Q) \geq 0$. Notice that $P = Q \implies D(P, Q) = 0$, but there may be $P \neq Q$ such that $D(P, Q) = 0$. However, if S is strictly proper, $D(P, Q) = 0 \iff P = Q$, which is the commonly used condition to define a statistical divergence (as for instance the Kullback-Leibler, or KL divergence). Therefore, each strictly proper scoring rule corresponds to a statistical divergence between probability distributions.

The energy score introduced in Sec. 1 is a strictly proper scoring rule for the class of probability measures P such that $\mathbb{E}_{X \sim P} \|X\|^\beta < \infty$ [35]. The related divergence is the square of the energy distance, which is a metric between probability distributions (72; see Appendix D.1)[2]. We will fix $\beta = 1$ in the rest of this work and we will write S_{E} in place of $S_{\mathrm{E}}^{(1)}$. Analogously, the kernel score is proper for the class of probability distributions for which $\mathbb{E}[k(X, X')]$ is finite (by Theorem 4 in [35]). Additionally, it is strictly proper under conditions which ensure that the MMD is a metric for probability distributions on $\mathcal{X}$ (see Appendix D.2). These conditions are satisfied, among others, by the Gaussian kernel (which we will use in this work):

$$k(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\gamma^2}\right),$$

in which γ is a scalar bandwidth. The divergence corresponding to the kernel score is the squared Maximum Mean Discrepancy (MMD, 37) relative to the kernel k (see Appendix D.2).

[2]The probabilistic forecasting literature [35] use a different convention for the energy score and the subsequent kernel score, which amounts to multiplying our definitions by 1/2. We follow here the convention used in the statistical inference literature [72, 13, 60]

2.2. The scoring rule posterior

Consider now a set of independent and identically distributed observations $\mathbf{y_n} \in \mathcal{X}^n$ sampled from a distribution P_0. The SR posterior [34, 52] is obtained by setting $\ell(y, \theta) = S(P_\theta, y)$ in the general Bayes update in Eq. (1):

$$\pi_S(\theta|\mathbf{y_n}) \propto \pi(\theta) \exp\left\{-w \sum_{i=1}^{n} S(P_\theta, y_i)\right\}. \tag{2}$$

The standard Bayes posterior is recovered from Eq. (2) by setting $w = 1$ and $S(P_\theta, y) = -\log p(y|\theta)$. Such choice of S is called the *log score*, is strictly proper, and corresponds to the Kullback-Leibler (KL) divergence. With the same S, $w \neq 1$ yields the fractional posterior [39, 7].

In this work, we focus on the SR posterior obtained with the energy and kernel scores (Sec. 1), which can be estimated from samples and, as such, make it suitable for likelihood-free inference. The SR posterior was previously studied in [34], which fixed $w = 1$ and adjusted the parameter value so that the posterior has the same asymptotic covariance matrix as the frequentist minimum scoring rule estimator (see Sec. 2.3), and in [52], which considered a time-series setting in which the task is to learn about the parameter value which yields the best prediction, given the previous observations.

Remark 1 (**Bayesian additivity**). The posterior in Eq. (2) satisfies Bayesian additivity (also called coherence, 10): sequentially updating the belief with a set of observations does not depend on the order the observations are received.

Remark 2 (**Non-invariance to change of data coordinates**). The SR posterior is in general not invariant to change of the coordinates used for representing the observations. This is a property common to loss-based frequentist estimators and to the generalized posterior obtained from them [56]; see Appendix B for more details.

2.3. Properties of the SR posterior

The SR posterior in Eq. (2) is a form of generalized Bayesian posterior [10]. [58] obtains asymptotic normality results broadly applicable to generalized Bayesian posterior, which can be adapted to the SR posterior (see the thesis [61] for more details). Notice that, for the posterior to concentrate asymptotically, the minimizer of the expected SR needs to be unique. Well-specified model and strictly proper SR S imply a unique minimizer; however, there are cases where the minimizer is unique even when the SR is not strictly proper or the model is not well specified. Similar asymptotic normality results can be obtained for Bayesian Synthetic Likelihood [31] and Approximate Bayesian Computation [30]. Notice however the result in [58] cannot be applied to the target of an MCMC where the scoring rule $S(P_\theta, y)$ is replaced with a sample estimate $\hat{S}(\mathbf{x_m}^{(\theta)}, y)$ (Sec. 3). Studying the asymptotic normality of such a target would be of interest, but we

leave it for future work; see [32] for a study on this in the case of Approximate Bayesian Computation.

In contrast to asymptotic normality for the traditional Bayesian posterior (Section 4.1.2 in 33), which ensures that the asymptotic covariance matrix matches that of the empirical maximum likelihood estimator, the asymptotic covariance matrix of the SR posterior does not match that of the frequentist minimizer of the SR, implying that asymptotic credible sets do not in general have correct frequentist coverage, even for strictly proper SR and well-specified model. This contrasts with Bayesian synthetic likelihood and regression-adjusted Approximate Bayesian Computation, for which correct frequentist coverage can be achieved for the asymptotic posterior under some conditions on the summary statistics [31, 49]. We also point out how a promising recent method for calibrating Bayesian inference for misspecified models [29] could be leveraged to fix this mismatch for the SR posterior; we leave this to future work.

For the energy and kernel score posteriors, a finite-sample bound on the probability of deviation of the posterior expectation of the divergence from the minimum divergence achievable by the model can be obtained (see [61] for a statement), similarly to what was done in [56] for a generalised posterior based on the kernel Stein discrepancy. Such bound does not require the model to be well specified nor the minimizer of the divergence to be unique.

Finally, for the energy and kernel score posteriors, it is possible to show an outlier-robustness result analogous to that obtained in [56] for the SR posterior; see [61] for a complete derivation.

3. Sampling the scoring rule posterior for LFI

Computing the energy and kernel scores, provided the likelihood is available, requires solving a double expectation, which is challenging in practice. In the following, we will show how the availability of samples from simulator models allows us to get unbiased estimates of the energy and kernel scores. Further, for differentiable simulator models (for which derivative of the simulated data w.r.t. to the parameters are available) we can also obtain unbiased estimators of the gradient of the scoring rules considered here under some regularity conditions. These derivatives can be effortlessly computed using automatic differentiation libraries for most simulator models [3].

To sample approximately from the scoring rule posterior, we propose a pseudo-marginal Monte Carlo Markov chain (PM-MCMC) algorithm using estimators of scoring rules computed from samples of the simulator model. In addition, we propose using stochastic gradient Monte Carlo Markov chain (SG-MCMC) algorithms for differentiable simulator models. When applicable, SG-MCMC avoids two known drawbacks of PM-MCMC, namely the curse of dimensionality limiting its application to high-dimensional parameter spaces and the "sticky" behaviour resulting in poor mixing for concentrated targets.

[3]Exceptions include simulator models with thresholding involved in their simulation process or when the simulated data is discrete.

3.1. Pseudo-marginal MCMC

Our PM-MCMC algorithm depends upon the existence of an estimate $\hat{S}(\mathbf{x}_\mathbf{m}^{(\theta)}, y)$ of $S(P_\theta, y)$, where $\mathbf{x}_\mathbf{m}^{(\theta)} = \{x_j^{(\theta)}\}_{j=1}^m$ is a set of samples $x_j^{(\theta)} \sim P_\theta$, and $\hat{S}$ is such that $\hat{S}(\mathbf{X}_\mathbf{m}^{(\theta)}, y) \to S(P_\theta, y)$ in probability as $m \to \infty$ (i.e., it estimates the SR consistently). Unbiased estimates for $S_\mathrm{E}^{(\beta)}$ and S_k can be obtained by unbiasedly estimating the expectations using samples $\mathbf{x}_\mathbf{m}^{(\theta)}$ as follows:

$$\hat{S}_\mathrm{E}^{(\beta)}(\mathbf{x}_\mathbf{m}^{(\theta)}, y) = \frac{2}{m}\sum_{j=1}^m \left\| x_j^{(\theta)} - y \right\|_2^\beta - \frac{1}{m(m-1)} \sum_{\substack{j,k=1 \\ k\neq j}}^m \left\| x_j^{(\theta)} - x_k^{(\theta)} \right\|_2^\beta;$$

$$\hat{S}_k(\mathbf{x}_\mathbf{m}^{(\theta)}, y) = \frac{1}{m(m-1)} \sum_{\substack{j,k=1 \\ k\neq j}}^m k(x_j^{(\theta)}, x_k^{(\theta)}) - \frac{2}{m}\sum_{j=1}^m k(x_j^{(\theta)}, y).$$

Notice how the above estimates can be negative; however, this is not an issue when employing these in our approximate MCMC methods targeting the SR posterior, as the above estimates are passed through the exponential defining the SR posterior, which makes the final target estimate positive.

For each proposed value of θ, we simulate $\mathbf{x}_\mathbf{m}^{(\theta)} = \{x_j^{(\theta)}\}_{j=1}^m$ and estimate the target in Eq. (2) with:

$$\pi(\theta) \exp\left\{ -w \sum_{i=1}^n \hat{S}(\mathbf{x}_\mathbf{m}^{(\theta)}, y_i) \right\}. \tag{3}$$

This procedure is an instance of pseudo-marginal MCMC [4], with target:

$$\pi_{\hat{S}}^{(m)}(\theta|\mathbf{y_n}) \propto \pi(\theta) p_{\hat{S}}^{(m)}(\mathbf{y_n}|\theta), \tag{4}$$

where:

$$p_{\hat{S}}^{(m)}(\mathbf{y_n}|\theta) = \mathbb{E}\left[\exp\left\{ -w \sum_{i=1}^n \hat{S}(\mathbf{X}_\mathbf{m}^{(\theta)}, y_i) \right\} \right].$$

For a single draw $\mathbf{x}_\mathbf{m}^{(\theta)}$, the quantity in Eq. (3) is in fact a non-negative and unbiased estimate of the target in Eq. (4); this approach is similar to what is proposed in [24] for inference with auxiliary likelihoods, which has also been used by [71] for BSL. As it was already the case for the latter, the target $\pi_{\hat{S}}^{(m)}(\theta|\mathbf{y_n})$ is not the same as $\pi_S(\theta|\mathbf{y_n})$ and depends on the number of simulations m; in fact, in general:

$$\mathbb{E}\left[\exp\left\{ -w \sum_{i=1}^n \hat{S}(\mathbf{X}_\mathbf{m}^{(\theta)}, y_i) \right\} \right] \neq \exp\left\{ -w \sum_{i=1}^n S(P_\theta, y_i) \right\},$$

even if $\hat{S}(\mathbf{x}_\mathbf{m}^{(\theta)}, y)$ is an unbiased estimate of $S(P_\theta, y)$. However, it is possible to show that, as $m \to \infty$, $\pi_{\hat{S}}^{(m)}$ converges to π_S:

Theorem 3.1. *Assume the following:*

1. $\hat{S}(\mathbf{X}_\mathbf{m}^{(\theta)}, y_i)$ *converges in probability to* $S(P_\theta, y_i)$ *as* $m \to \infty$ *for all* $i = 1, \dots, n$.
2. $\sup_m \mathbb{E}\left[\exp\{-(1+\delta)w \sum_{i=1}^n \hat{S}(\mathbf{X}_\mathbf{m}^{(\theta)}, y_i)\}\right] < \infty$ *for some* $\delta > 0$.
3. $\inf_m \int_\Theta p_{\hat{S}}^{(m)}(\mathbf{y_n}|\theta)\pi(\theta)d\theta > 0$ *and* $\sup_m p_{\hat{S}}^{(m)}(\mathbf{y_n}|\theta) \leq g(\theta)$ *where* $\int_\Theta g(\theta)\pi(\theta)d\theta < \infty$.

Then,

$$\lim_{m\to\infty} \pi_{\hat{S}}^{(m)}(\theta|\mathbf{y_n}) = \pi_S(\theta|\mathbf{y_n}).$$

The above result (proven in appendix A) is an extension of the one in [24] for Bayesian inference with an auxiliary likelihood. If $\hat{S}(\mathbf{x}_\mathbf{m}^{(\theta)}, y_i) \geq S(\hat{P}_\theta, y_i)$, where $\hat{P}_\theta$ is the empirical distribution determined by $\mathbf{x}_\mathbf{m}^{(\theta)}$, and S is a proper scoring rule, Assumption 2 above is automatically verified (see Appendix A.1); however, notice that, the scoring rule estimates $\hat{S}_k$ and $\hat{S}_E$ for the kernel and energy scores are not lower bounded by the scoring rules of the empirical distribution.

The assumptions and conclusions are stated considering a fixed value of $\mathbf{y_n}$; if the assumptions were to hold almost surely over $\mathbf{Y_n}$, the conclusion would also hold almost surely.

In practice, in place of the vanilla pseudo-marginal approach discussed above, we use a correlated pseudo-marginal MCMC [17, 20, 68], which reuses the random numbers used in model simulations over subsequent proposed parameter values. This correlates the target estimates at subsequent steps and reduces the chances of the chain getting stuck due to atypical random number draws. Specifically, the m simulations used in the posterior estimate (Eq. 3) are split into G groups; at each MCMC step, a new set of random numbers is proposed for the simulations in a randomly chosen group (alongside the proposed value for θ), and accepted or rejected in the standard way. This algorithm still targets Eq. (4).

3.2. Stochastic gradient MCMC

For the scoring rules used across this work, as well as any weighted sum of those, we can write $S(P_\theta, y) = \mathop{\mathbb{E}}_{X,X'\sim P_\theta} g(X, X', y)$ for some function g; namely, the SR is defined through an expectation over (possibly multiple) samples from P_θ. In the following, we assume random samples from the simulator model P_θ can be written as $X = h_\theta(Z)$ where Z follows a base distribution Q independent of the parameters θ. Now:

$$\begin{aligned}
\nabla_\theta S(P_\theta, y) &= \nabla_\theta \mathop{\mathbb{E}}_{X,X'\sim P_\theta} g(X, X', y) \\
&= \nabla_\theta \mathop{\mathbb{E}}_{Z,Z'\sim Q} g(h_\theta(Z), h_\theta(Z'), y) \\
&= \mathop{\mathbb{E}}_{Z,Z'\sim Q} \nabla_\theta g(h_\theta(Z), h_\theta(Z'), y).
\end{aligned}$$

In the latter equality, the exchange of derivative and expectation is valid if both g and h_θ are differentiable. Moreover, it is also valid with some non-differentiable h_θ; for instance, Theorem 5 in [9] ensures this for h_θ being a neural network with Lipschitz layers (which is the case for the vast majority of neural networks used in practice), provided Q satisfies a mild moment condition and g is continuously differentiable and satisfies two growth conditions. This guarantees our method is applicable to the model we study in Sec. 4.3.3, which is defined by a neural network.

Based on the above equality, we estimate the gradient of the scoring rule as follows:

$$\widehat{\nabla_\theta} S(P_\theta, y) = \frac{1}{m(m-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \nabla_\theta g\left(h_\theta(Z_i), h_\theta(Z'_j), y\right), \quad Z_i \perp\!\!\!\perp Z'_j \sim Q.$$

In practice, this can easily be obtained by implementing the function h_θ using automatic-differentiation libraries [66].

By relying on this construction, we adapt two existing SG-MCMC [59] algorithms (stochastic gradient Noose-Hoover thermostat [22] and Preconditioned Stochastic Gradient Langevin [48]) to sample from the scoring rule posterior. As mentioned above, these algorithms are approximate, but the computational advantage they provide overweights the induced approximation.

Alternatively, Piecewise-Deterministic Markov Processes (PDMP, 27) allow exact sampling with an unbiased estimate of the log-target gradient; unfortunately, however, the exact implementation of the existing algorithms requires computing an upper bound of the log-target gradient which is intractable for most practical use cases. To avoid this, approximate methods [64, 15] are developed, which are however inconvenient for general target distributions compared to SG-MCMC methods.

Adaptive Stochastic Gradient Langevin Dynamics (adSGLD) The earliest known stochastic gradient MCMC algorithm [78] is based upon the (Overdamped) Langevin Diffusion, defined by the following Stochastic Differential Equation:

$$\mathrm{d}\theta(t) = -\frac{1}{2}\nabla_\theta U(\theta(t))\mathrm{d}t + \mathrm{d}B_t.$$

For the SR posterior, $U(\theta) = \log \pi(\theta) - w \sum_{i=1}^{n} S(P_\theta, y_i)$, $\theta \in \mathbb{R}^d$ and $B_t \in \mathbb{R}^d$ is standard Brownian Motion. Under suitable regularity conditions, this continuous-time diffusion has $\pi_S(\theta|\mathbf{y_n})$ as its stationary distribution [73, 69]. In practice, we are unable to simulate this stochastic process exactly. Hence, numerical integration schemes are used to generate samples. For instance, the Euler-Maruyama method consists of the following update:

$$\theta_{t+1} \leftarrow \theta_t - \frac{\epsilon}{2}\nabla_\theta U(\theta(t)) + \sqrt{\epsilon} Z$$

repeated over t, where Z is a d-dimensional standard normal random vector and ϵ is a discretisation step size. Following [78], we propose to use the unbiased

estimate of the gradient of $\nabla_\theta U(\theta(t))$,

$$\widehat{\nabla_\theta} U(\theta) = \nabla_\theta \log \pi(\theta) - w \sum_{i=1}^{n} \widehat{\nabla_\theta} S(P_\theta, y_i)$$

in the above update equation; this method is called Stochastic Gradient Langevin Dynamics (SGLD). Using a sequence $\{\epsilon_i\}_{i=1}^N$ converging to 0 and taking $m \to \infty$, under some condition, [78] shows that SGLD samples from the scoring rule posterior.

In practice, however, we do not have $\epsilon_i \to 0$ or $m \to \infty$. Hence, to ensure sampling with minimal bias for our noisy gradient scenario, we utilize the adaptive Langevin dynamics originally proposed in [43] and later used for Bayesian inference in [22]. We will refer to this algorithm as adaptive stochastic gradient Langevin dynamics (adSGLD), which runs on an augmented space (θ, p, ξ), where θ represents the parameter of interest, $p \in \mathbb{R}^d$ represents the momentum and ξ represents an adaptive thermostat controlling the mean kinetic energy $\frac{1}{n}\mathbb{E}[p^\top p]$, along with a diffusion factor $\mathcal{A}$. Thus, the new dynamics are as follows:

$$\begin{cases} \mathrm{d}\theta_t = p_t \ \mathrm{d}t \\ \mathrm{d}p_t = -\nabla_\theta U(\theta(t))\mathrm{d}t - \xi p_t \ \mathrm{d}t + \sqrt{2\mathcal{A}}\mathcal{N}(0, I\mathrm{d}t) \\ d\xi = \left(\frac{1}{n} p_t^\top p_t - 1\right) dt. \end{cases}$$

Theoretical properties and convergence of adSGLD algorithm have been studied in [22], [47] and [46]. Below, we state the adSGLD algorithm, which requires fixing the hyperparameters ϵ (step size) and $\mathcal{A}$.

Algorithm 1 adSGLD Algorithm for scoring rule posterior

Input: $\mathcal{A}$, ϵ, θ_0, N
Output: $\{\theta_i\}_{i=1}^N$ **samples**

Initialise $P_0 \sim N(0, I)$ and $\xi_0 \leftarrow \mathcal{A}$
for $i = 1$ to N **do**:
  Estimate $\widehat{\nabla}_\theta U(\theta_{i-1})$
  $P_i \leftarrow P_{i-1} - \xi_{i-1} P_{i-1}\epsilon - \widehat{\nabla}_\theta U(\theta_{i-1})\epsilon + \sqrt{2\mathcal{A}}N(0, \epsilon)$
  $\theta_i \leftarrow \theta_{i-1} + P_i\epsilon$
  $\xi_i \leftarrow \xi_{i-1} + \left(\frac{1}{n} P_i^\top P_i - 1\right)\epsilon$
end for

Preconditioned Stochastic Gradient Langevin Dynamics (pSGLD, 48) This algorithm preconditions the log-target with a diagonal matrix $G(\theta)$ obtained through a running average of the squared gradients using the following update equations:

$$G(\theta_{t+1}) = \mathrm{diag}\left(\mathbf{1} \oslash \left(\lambda \mathbf{1} + \sqrt{V(\theta_{t+1})}\right)\right)$$
$$V(\theta_{t+1}) = \alpha V(\theta_t) + (1-\alpha)\widehat{\nabla}_\theta U(\theta_t) \odot \widehat{\nabla}_\theta U(\theta_t)$$

with $\oslash$ and $\odot$ denoting element-wise matrix division and product respectively. The hyperparameter λ is a small bias term to avoid the degeneration of the preconditioner, while $\alpha \in (0,1)$ is a relative weighting between the previous and current gradients. This algorithm performs well for non-convex posteriors on high-dimensional space, and in particular for the complicated posteriors characterized by deep neural networks. We state the algorithm for pSGLD below.

Algorithm 2 pSGLD Algorithm for scoring rule posterior

Input: λ, α, ϵ, θ_0, N
Output: $\{\theta_i\}_{i=1}^N$ **samples**

1: Initialise $V_0 \leftarrow \mathbf{0}$
2: **for** $i = 1$ to N **do**:
3: Estimate $\widehat{\nabla}_\theta U(\theta_i)$
4: $V(\theta_i) \leftarrow \alpha V(\theta_{i-1}) + (1-\alpha)\widehat{\nabla}_\theta U(\theta_i) \odot \widehat{\nabla}_\theta U(\theta_i)$
5: $G(\theta_i) \leftarrow \text{diag}\left(\mathbf{1}\left(\lambda \mathbf{1} + \sqrt{V(\theta_i)}\right)\right)$
6: $\theta_{i+1} \leftarrow \theta_i + \frac{\epsilon}{2} G(\theta_i) U(\theta_i) + \mathcal{N}(0, \epsilon G(\theta_i))$
7: **end for**

In practice, we set λ to 10^{-5} and α to 0.99.

Choice of step size ϵ For SG-MCMC algorithms, choosing the step-size ϵ is critical, as it represents a trade-off between the speed of convergence or mixing performance and the discretisation error. In practice, SG-MCMC algorithms are often used with a constant step size due to slow mixing when $\epsilon \approx 0$. To tune ϵ, we use a modified version of the multi-armed bandit algorithm based on the kernelized Stein discrepancy proposed in [16]. This algorithm identifies each arm with a specific hyperparameter configuration, and for a fixed time budget, sequentially eliminates poor hyperparameter configurations based on the kernelized Stein discrepancy between the samples and the target distribution.

4. Empirical studies

4.1. Comparison between PM-MCMC and SG-MCMC

To compare PM-MCMC and SG-MCMC (specifically, the adSGLD algorithm), we perform an empirical study on the univariate g-and-k model [70]. The latter is defined in terms of the inverse of its cumulative distribution function F^{-1}. Given a quantile q, we define:

$$F^{-1}(q) = A + B\left[q + 0.8\frac{1-e^{-gz(q)}}{1+e^{-gz(q)}}\right]\left(1+z(q)^2\right)^k z(q), \tag{5}$$

where the parameters A, B, g, k are broadly associated with the location, scale, skewness and kurtosis of the distribution, and $z(q)$ denotes the q-th quantile of the standard normal distribution $\mathcal{N}(0,1)$. Likelihood evaluation for this model is costly as it requires numerical inversion of F^{-1}; instead, sampling is immediate by drawing $z \sim \mathcal{N}(0,1)$ and inputting it in place of $z(q)$ in the expression above.

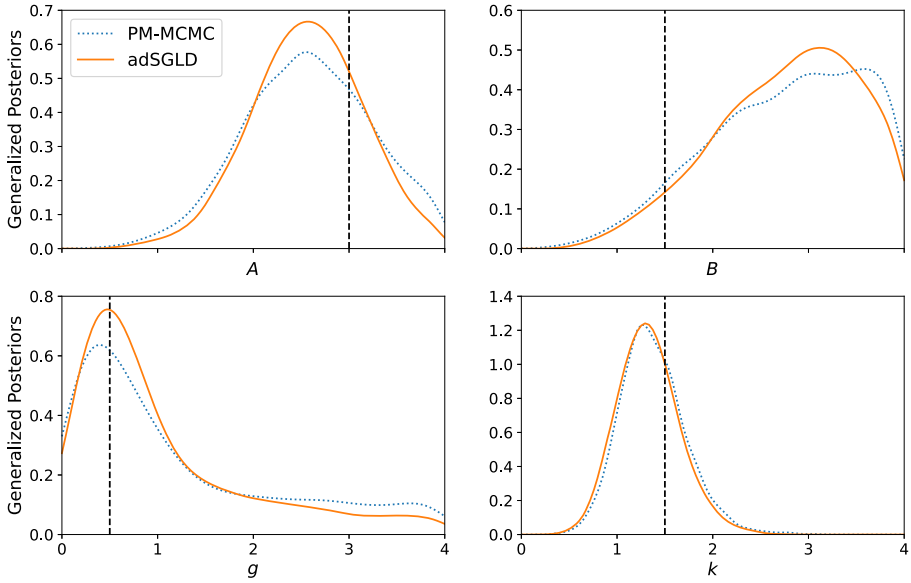


FIG 1. ***Comparison of adSGLD and PM-MCMC*** *to sample from the marginals of the energy score posterior for the g-and-k model obtained with* $n = 10$*. Vertical lines denote true parameter values. For both, 100000 samples with 10000 burn-in were used.*

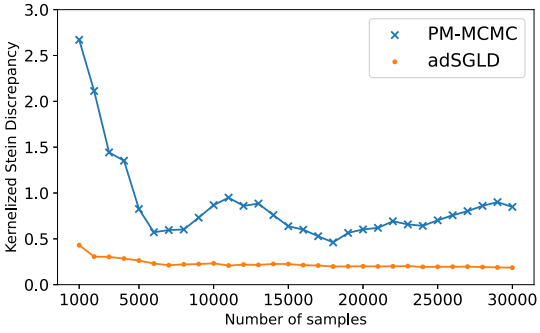
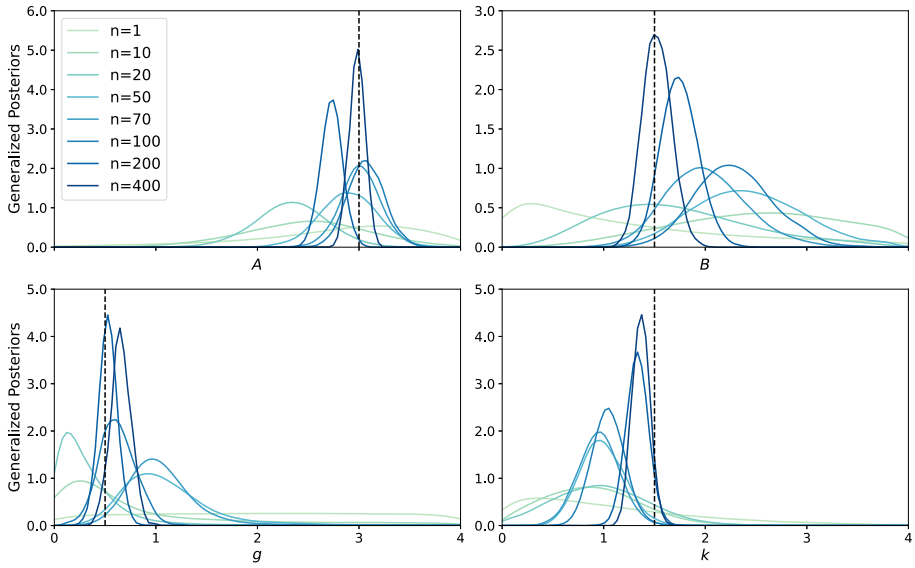
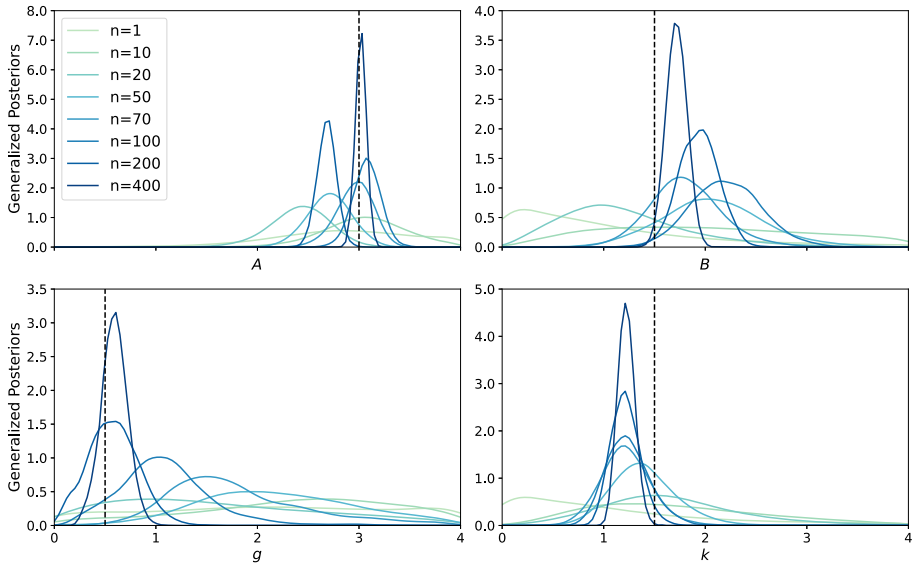


FIG 2. ***Kernelized Stein Discrepancy*** *for first 30000 MCMC samples for the energy score posterior for the g-and-k model, on* $n = 10$ *observations, sampled with adSGLD and PM-MCMC. KSD uses the inverse multi-quadratic kernel, with the gradients estimated using 500 simulated observations from each parameter. adSGLD both converges faster and is more accurate than the PM-MCMC algorithm.*



(a) Marginals of energy score posterior.



(b) Marginals of kernel score posterior.

FIG 3. ***Posterior concentration of univariate g-and-k model,*** *illustrated by marginals of (a) energy score and (b) kernel score posteriors for the different parameters of the univariate g-and-k model, with increasing number of observations* ($n = 1, 10, 20, \ldots, 400$). *Darker (respectively lighter) colours denote a larger (smaller) number of observations. The densities are obtained by kernel density estimator on the MCMC output. The energy and kernel score posteriors concentrate around the true parameter value (dashed vertical line).*

We use uniform priors on $[0,4]^4$ on the sets of parameters $\theta = (A, B, g, k)$. For $n = 10$ observations from true parameter values $A^\star = 3$, $B^\star = 1.5$, $g^\star = 0.5$, $k^\star = 1.5$, we perform inference with the energy score Posterior with $w = 1$, setting the number of simulations per parameter value to $m = 500$ and run adSGLD and PM-MCMC for 110000 steps. Additional experimental details are reported in Appendix F.1.1.

Figure 1 shows a kernel density estimate of the samples obtained with the two methods: the two densities are similar, with the PM-MCMC one slightly broader. As both sampling methods are asymptotically biased, we cannot rely on traditional MCMC diagnostics (such as the R-hat and the autocorrelation function) to quantitatively evaluate sample quality, as those only evaluate properties of the chain itself and are thus unable to measure the discrepancy between samples from an approximate sampler and exact target. To this aim, we employ the kernelized Stein discrepancy (KSD) proposed in [36] which, conveniently, can be estimated by using MCMC samples and unbiased estimates of the gradient of the log target (see Appendix C). We compute the KSD with an increasing number of samples obtained from the two methods, thus allowing us to investigate which algorithm converges faster. The results can be seen in Fig. 2: the adSGLD algorithm converges faster than the PM-MCMC algorithm and asymptotically produces samples that are a better approximation to the target distribution. Based on the superior performance of the adSGLD algorithm here, we will employ it for sampling from the SR posterior in the remaining simulation studies as all our considered simulator models are differentiable unless otherwise specified. For comparison, results with PM-MCMC for some of the setups considered in the main body of the paper are reported in Appendix G.

4.2. Posterior concentration of univariate g-and-k model

To empirically evaluate the concentration of the SR posterior, we consider the g-and-k model introduced in Sec. 4.1 and sample from the energy and kernel score posteriors for an increasing number of observations (up to $n = 400$) generated from $A^\star = 3$, $B^\star = 1.5$, $g^\star = 0.5$, $k^\star = 1.5$.

For the same value of w, the scale of the two SR posteriors is different as it depends on the values taken by the SR itself. As here we aim to compare the concentration speed of the two posteriors, we set w such that they have roughly the same scale (for the same number of observations. In other use cases, as mentioned in the introduction, w can be selected to achieve different goals (often, to match some frequentist property, see 54, 76, 57).

In practice, we adapt a method proposed in [10] which does not require repeated posterior inference and knowledge of the likelihood function. Specifically, notice that:

$$\log \underbrace{\left\{ \frac{\pi_S(\theta|y)}{\pi_S(\theta'|y)} \Big/ \frac{\pi(\theta)}{\pi(\theta')} \right\}}_{\mathrm{BF}_S(\theta,\theta';y)} = -w \left\{ S(P_\theta, y) - S(P_{\theta'}, y) \right\}$$

where $\mathrm{BF}_S(\theta, \theta'; y)$ denotes the Bayes Factor of θ with respect to θ' for observation y. Therefore, w can be determined by fixing $\mathrm{BF}_S(\theta, \theta'; y)$ for a single choice of θ, θ', y. Consider now another SR posterior $\pi_{S'}(\theta|y)$ with Bayes Factor $\mathrm{BF}_{S'}$; setting:

$$w = -\frac{\log \mathrm{BF}_{S'}(\theta, \theta'; y)}{S(P_\theta, y) - S(P_{\theta'}, y)},$$

ensures $\mathrm{BF}_{S'}(\theta, \theta'; y) = \mathrm{BF}_S(\theta, \theta'; y)$. If π_S and $\pi_{S'}$ are obtained from the same prior distribution and the latter uses $w = 1$, that corresponds to

$$w\{S(P_\theta, y) - S(P_{\theta'}, y)\} = S'(P_\theta, y) - S'(P_{\theta'}, y).$$

As we have no reason to prefer a specific choice of (θ, θ'), we set w to be the median of $\frac{S'(P_\theta, y) - S'(P_{\theta'}, y)}{S(P_\theta, y) - S(P_{\theta'}, y)}$ over values of (θ, θ') sampled from the prior. In doing so, we ensure the median variation of the SR (multiplied by the corresponding w between two parameter values sampled from the prior is the same across the two posteriors. Additionally, if P_θ is an intractable-likelihood model, we estimate $S(P_\theta, y)$ and $S'(P_\theta, y)$ by generating data $\mathbf{x}_\mathbf{m}^{(\theta)}$ for each considered values of θ.

Hence, we set $w = 1$ for the energy score posterior and use the above method to tune w for the kernel score posterior, yielding $w = 28.1$; the bandwidth of the Gaussian kernel was tuned as discussed in Appendix E. Additional experimental details are reported in Appendix F.1.1. Figure 3 reports the results; with the chosen values of w, the two posteriors concentrate at roughly the same speed close to the true parameter values.

In Appendix G.1 we report similar results achieved with PM-MCMC; due to the stickyness of the chain, those only run satisfactorily up to $n = 100$.

4.3. Comparison with popular LFI methods

We present here simulation studies to compare our approach with two popular LFI schemes, Bayesian Synthetic Likelihood (BSL, 71) and Approximate Bayesian Computation (ABC, 50), and showcase the ability of SG-MCMC to sample from the scoring rule posterior of models with high-dimensional parameter space.

In general, the performance of both ABC and BSL depends on the choice of a set of summary statistics, which makes the method approximate (unless the chosen statistics are sufficient, which is seldom the case). On the contrary, our proposed methodology directly computes the SRs on the raw data and, as such, avoids the need to determine suitable statistics. Nevertheless, to provide a fair comparison, we tested BSL and ABC both on the raw data and on summary statistics, as detailed below. Even so, our method still performs better than ABC and BSL.

We first study the posterior concentration of the energy and kernel score posteriors compared to BSL in Sec. 4.3.1 for both well-specified and misspecified models; next, in Sec. 4.3.2, we consider a meteorological model with high-dimensional time-series dataset and compare the posterior predictive accuracy

of the scoring rule posterior with that obtained with SMC-ABC [19]. Finally in Sec. 4.3.3, we consider a neural extension of the meteorological model considered in Sec. 4.3.2 with a high-dimensional (> 100) parameter space; there, SG-MCMC allows us to sample from the high-dimensional SR posterior, thus enabling a better posterior predictive accuracy than the lower dimensional model considered in Sec. 4.3.2.

Throughout, the kernel score uses the Gaussian kernel with bandwidth set from simulations as illustrated in Appendix E; further, we set $w = 1$ in the energy score posterior and set w for the kernel score posterior with the strategy discussed in Sec. 4.2. Unless specified otherwise, the LFI techniques are run using the `ABCpy` Python library [26], code for reproducing all results is available as Supplementary Material [63].

4.3.1. Comparison with Bayesian synthetic likelihood: multivariate g-and-k model

Bayesian Synthetic Likelihood (BSL, 71) considers the following approximate posterior:

$$\pi_{\mathrm{SL}}(\theta|s(\mathbf{y_n})) \propto \pi(\theta)\mathcal{N}(s(\mathbf{y_n}); \mu_{s,\theta}, \Sigma_{s,\theta}), \tag{6}$$

where s is a set of summary statistics, $\mu_{s,\theta}$ and $\Sigma_{s,\theta}$ represent the mean and variance matrix of s at θ and $\mathcal{N}(\cdot; \mu_{s,\theta}, \Sigma_{s,\theta})$ denotes the multivariate normal density with mean vector $\mu_{s,\theta}$ and variance matrix $\Sigma_{s,\theta}$. BSL typically employs a PM-MCMC where multiple simulated datasets $\mathbf{x}_{\mathbf{m}}^{(\theta)}$ at each θ value are used to estimate $\mu_{s,\theta}$ and $\Sigma_{s,\theta}$ [71], analogously to what we discussed in Sec. 3. If suitable summaries are chosen (for instance, the average), the central limit theorem ensures that the summaries are asymptotically normal. However, the formulation in Eq. (6) does not satisfy Bayesian additivity [10].

We consider here the multivariate extension [23, 42] of the univariate g-and-k model introduced earlier. Specifically, we draw a multivariate normal $(Z^1, \ldots, Z^5) \sim \mathcal{N}(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{5\times5}$ has a sparse correlation structure: $\Sigma_{kk} = 1$, $\Sigma_{kl} = \rho$ for $|k - l| = 1$ and 0 otherwise; each component of Z is then transformed as in the univariate case (Eq. (5)). The sets of parameters are $\theta = (A, B, g, k, \rho)$. We use uniform priors on $[0, 4]^4 \times [-\sqrt{3}/3, \sqrt{3}/3]$.

For BSL, we adapt the summary statistics from [3], which studied a 3-dimensional g-and-k model with different marginal parameters for each dimension and used 4 marginal statistics for each dimension and cross-covariance estimates for each pair of dimensions. In our case, we pool together all dimensions of each simulated dataset to compute the 4 marginal statistics and we compute the 10 possible pairs of dimensions cross covariance statistics. This leads to a total of 14 statistics.

We also attempt using BSL directly on the raw data by considering a separate likelihood term for each observation y_i:

$$\pi_{\mathrm{SL}}(\theta|\mathbf{y_n}) \propto \pi(\theta) \prod_{i=1}^{n} \mathcal{N}(y_i; \mu_\theta, \Sigma_\theta), \tag{7}$$

where $\mu_\theta, \Sigma_\theta$ are estimates for the mean and variance matrix of $x \sim P_\theta$; notice how the posterior in Eq. (7) is a specific case of our SR posterior (Eq. 2) for $w = 1$ and the so-called *Dawid–Sebastiani* scoring rule (Appendix D.3), which is non-strictly proper (hence, multiple minimizers of the expected score can exist even for well-specified models, which implies that the posterior may fail to concentrate asymptotically).

For the SR posteriors, we use adSGLD with $m = 500$ and with 110000 steps and 10000 burn-in; additional experimental details for the SR posteriors are given in Appendix F.1.2. For BSL without summaries, we use correlated PM-MCMC with $m = 500$, $G = 500$ and run for 110000 steps, of which 10000 are burned in (where G refers to the number of groups in correlated PM-MCMC, see Section 3.1). Finally, recall how BSL with summaries, at each MCMC step, requires simulating M datasets with n simulations each [71], so that the overall number of simulations is $m = M \cdot n$. Each of the M datasets is used to estimate one set of statistics values, and those M values are in turn used to estimate $\mu_{s,\theta}$ and $\Sigma_{s,\theta}$. Thus, we run MCMC with 5500 steps, of which 1000 are burned in and used to estimate the proposal covariance matrix for the remainder of the chain. At each MCMC step, a fixed number of $m = 20,000$ simulations are generated and grouped into M datasets, with M decreasing linearly when n increases. This choice ensures that the total number of simulations used by BSL with summaries is equal to that used by the SR posterior[4] and is identical for each value of n, while also making sure that M is still large enough to give a reliable estimate of $\mu_{s,\theta}$ and $\Sigma_{s,\theta}$ even for the largest value of n we tried[5]. BSL with summaries is run using the `R` package `bsl` [2], which includes the summary statistics discussed above.

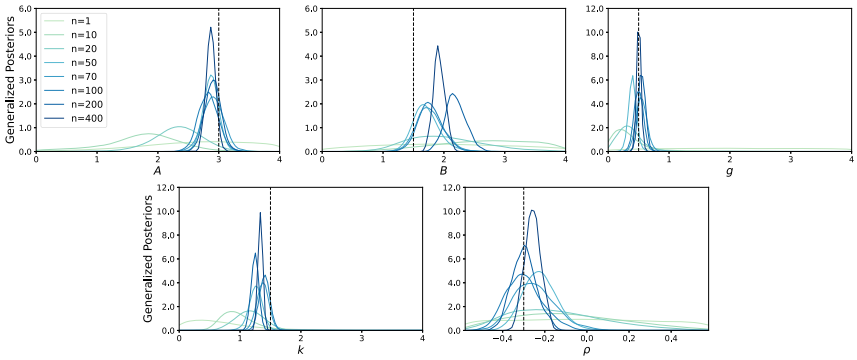
In Appendix G, results obtained using PM-MCMC for the SR posteriors are provided. The same appendix provides results for BSL without summaries on the univariate g-and-k model; there, PM-MCMC runs satisfactorily up to $n = 100$, showing how it fails to concentrate as it is based on non-strictly proper SRs.

Well-specified case We consider an increasing number of synthetic observations generated from parameter values $A^\star = 3$, $B^\star = 1.5$, $g^\star = 0.5$, $k^\star = 1.5$ and $\rho^\star = -0.3$, up to $n = 400$. The results are given in Figure 4. With increasing n, both the energy and kernel score posterior concentrate close to the true value for all parameters (dashed vertical line), as expected when using strictly proper SRs. For this example, the PM-MCMC targeting the BSL posteriors with and without summaries do not converge beyond 50 and 10 observations respectively. Moreover, BSL with summaries for $n = 50$ is highly concentrated far from the true parameter values.

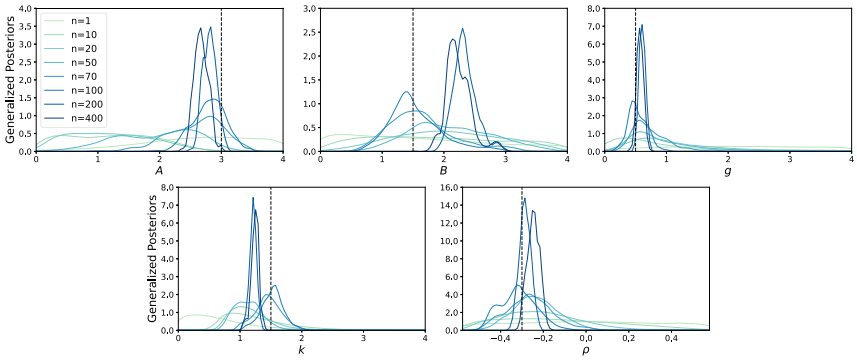
Misspecified setup Next, we consider as data generating process the Cauchy distribution, which has fatter tails than the g-and-k one. The five components of

[4] For this, we double the number of simulations to calculate the total budget, to take into account the gradient computation, which is absent in BSL.

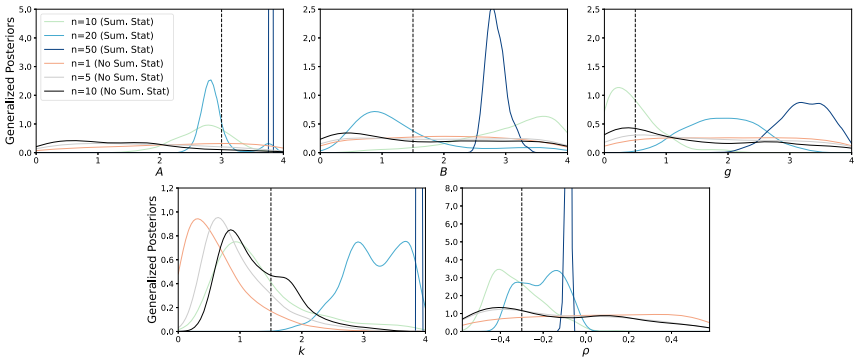
[5] Another option was that of reducing the number of MCMC steps as n increases, but that impacts the convergence of the chain and renders comparison even harder.



(a) Marginals of energy score posterior

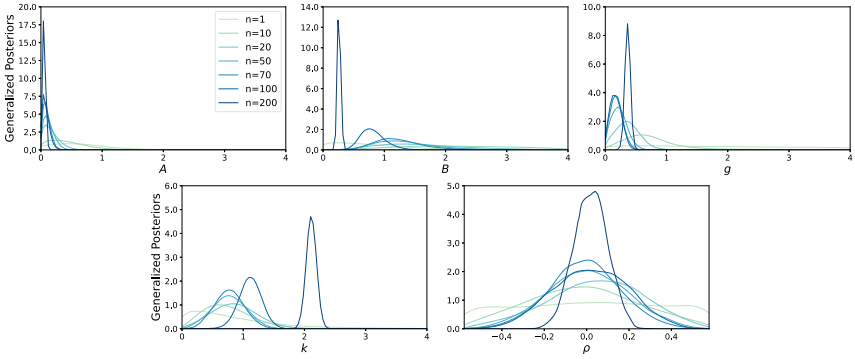


(b) Marginals of kernel score posterior

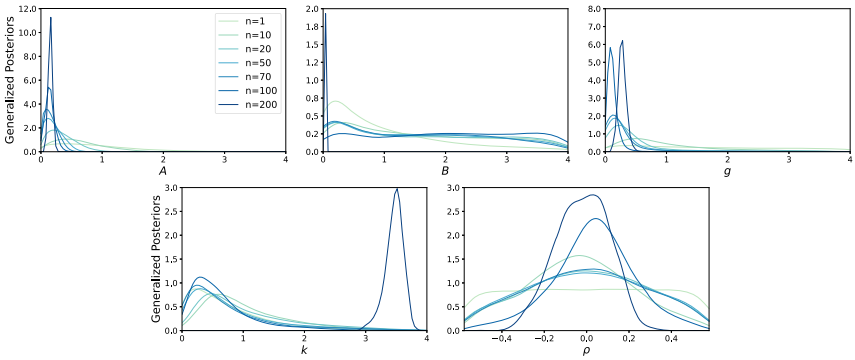


(c) Marginals of Bayesian synthetic likelihood (with and without summary statistics)

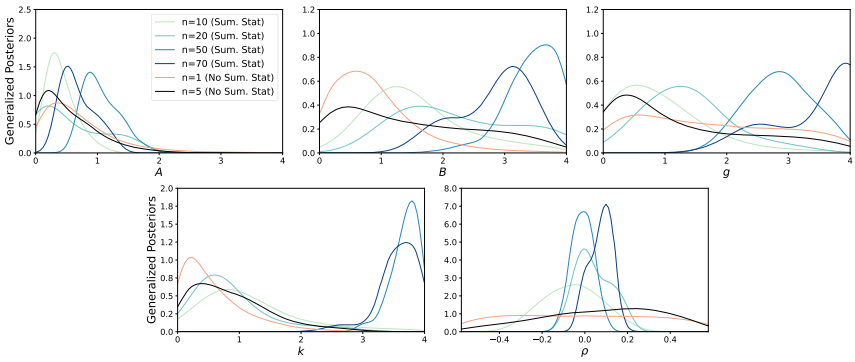
FIG 4. ***Posterior concentration of well-specified multivariate g-and-k model,*** *illustrated by marginals of (a) energy score, (b) kernel score and (c) Bayesian synthetic likelihood posteriors, with increasing number of observations* ($n = 1, 10, \ldots, 400$). *Darker (respectively lighter) colours denote a larger (smaller) number of observations. The vertical line represents the true parameter value. Both the energy and kernel score posteriors (run with adSGLD) concentrate close to the true parameter value, while the PM-MCMC targeting the BSL posteriors with and without summaries do not converge beyond 50 and 10 observations respectively.*



(a) Marginals of energy score posterior

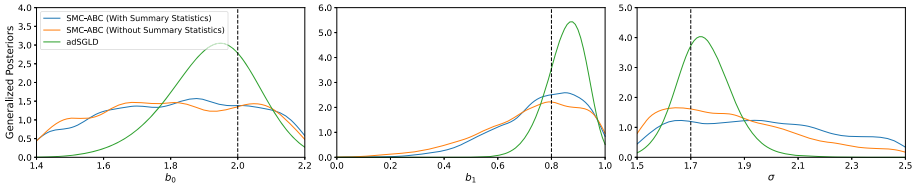


(b) Marginals of kernel score posterior

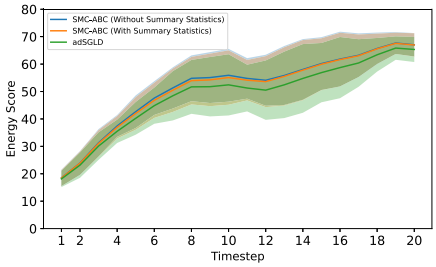


(c) Marginals of Bayesian synthetic likelihood (with and without summary statistics)

FIG 5. ***Posterior concentration of misspecified multivariate g-and-k model,*** *illustrated by marginals of (a) energy score, (b) kernel score and (c) Bayesian synthetic likelihood posteriors, with increasing number of observations ($n = 1, 10, \ldots, 400$). Darker (respectively lighter) colours denote a larger (smaller) number of observations. The vertical line represents the true parameter value. Both the energy and kernel score posteriors (run with adSGLD) concentrate close to the true parameter value, while the PM-MCMC targeting the BSL posteriors with and without summaries do not converge beyond 50 and 10 observations respectively*



(a) Marginal Posteriors.



(b) Predictive accuracy using energy score.

FIG 6. ***Comparison between SMC-ABC and energy score posteriors inferred using 250,000 model simulations, for the linearly parametrized Lorenz96 model.*** *(a) Marginal posterior distribution of the parameters of SMC-ABC posterior and energy score posterior using adSGLD, for a single observed set* $\mathbf{x_n}$ *(vertical line representing the true parameter* $\theta^\star$*). (b) Energy score between posterior predictive and each time-step of the original observation. This is repeated for 5 observations* $\mathbf{x_n}$ *(each using* $n = 10$ *here), and a t-distribution at each time-step is fitted to the energy score values. The solid line and shaded region respectively represent the mean and the 95% confidence interval of the fitted t-distribution. A lower energy score indicates better predictive performance.*

each observation are drawn independently from the univariate Cauchy distribution (i.e., no correlation between components). For the SR posteriors, we use the values of w which were obtained with our heuristics in the well-specified case; additional experimental details are reported in Appendix F.1.2. Results are in Figure 5. We consider an increasing number of observations n up to 200. The energy and kernel score posteriors concentrate on slightly different parameter values, corresponding to the unique minimizers of the expected SR (which are therefore different in these two cases). The PM-MCMC targeting the BSL posteriors with and without summaries do not converge beyond 70 and 5 observations respectively. This shows how the amenability of the SR posterior to SG-MCMC (which is instead not the case for BSL) is crucial for good performance.

4.3.2. Comparison with approximate Bayesian computation: stochastic Lorenz96 model

The Lorenz96 model [53] is an important benchmark in meteorology [5] and was previously studied in the LFI literature [77, 40, 62]. Here, we consider the stochastic *parametrized* version introduced by [79], defined by the following set

of Ordinary Differential Equations (ODEs):

$$\frac{dx_k}{dt} = -x_{k-1}(x_{k-2} - x_{k+1}) - x_k + 10 - g(x_k, t; \theta); \quad k = 1, \ldots, K,$$

where cyclic boundary conditions imply that we take $K + 1 = 1$ in the indices. The stochastic forcing term g depends on parameters $\theta = (b_0, b_1, \sigma_e)$, and is defined upon discretizing the ODEs with a time-step Δt:

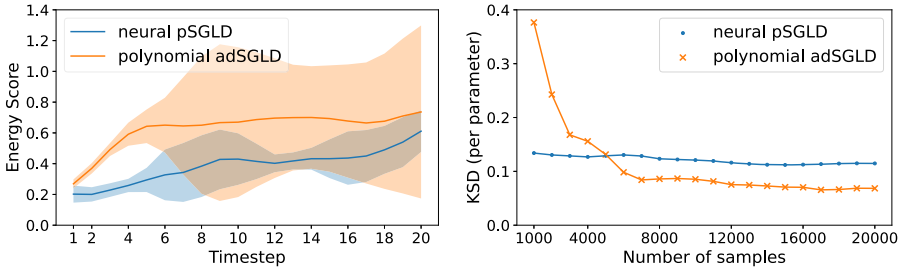
$$g(x, t; \theta) = b_0 + b_1 x + \sigma_e \eta(t), \quad \eta(t) \sim \mathcal{N}(0, 1). \tag{8}$$

In practice, we took $K = 8$ and integrated the model using the Euler-Maruyama scheme starting from a fixed initial condition $x(0)$ for 20 additional time-steps on the interval $t \in [0, 1.5]$ (corresponding to $\Delta t = 3/40$). We generate 5 independent sets of observed data $\mathbf{x_n}$, each using $n = 10$ time-series simulated from the model using $\theta^\star = (2, 0.8, 1.7)$. The integration output is an 8-dimensional time series with 20 time steps. As a prior distribution, we consider a uniform distribution on the region $[1.4, 2.2] \times [0, 1] \times [1.5, 2.5]$.

We run inference for the energy score posterior using adSGLD with $m = 10$ and 25000 MCMC steps, of which 5000 are burned-in. We compare the inferred energy score posterior with the posterior obtained by Sequential Monte Carlo Approximate Bayesian Computation (SMC-ABC, 19) using the Euclidean distance between either the full simulated and observed datasets or the summary statistics suggested in [38] (the temporal mean and variance of $x_k(t)$, the auto-covariance of $x_k(t)$ with time lag 1, and the covariance of $x_k(t)$ with its two neighbours $x_{k-1}(t)$ and $x_{k+1}(t)$; averaged over the index k, leading to a 6-dimensional set of statistics) as discrepancy measure. The SMC-ABC algorithm was run for 25 generations with $m = 10$ simulations for every parameter value to draw 1000 samples from the posterior distribution; with this setup, the algorithms each use $250,000$ model simulations. Further details are given in Appendix F.2. The comparison between these posteriors in Figure 6a illustrates how the energy score posterior assigns more probability to parameter values close to $\theta^\star$ than the SMC-ABC posteriors using the full data or the summary statistics. Moreover, to assess the out-of-sample performance of the inferred posterior, we implement the following posterior predictive check: given draws from a posterior $\pi(\theta|\mathbf{x_n})$, we generate simulations from the model for the corresponding parameter value, which are therefore samples from the posterior predictive

$$p(y_{\text{new}}|\mathbf{x_n}) = \int p(y_{\text{new}}|\theta)\pi(\theta|\mathbf{x_n})d\theta;$$

from these samples, we assess how well the posterior predictive matches the original observation by computing the energy score between the posterior predictive distribution and the observations $\mathbf{x_n}$ at each time step. The results in Figure 6b show how the energy score posterior predictive matches the original observation better than the SMC-ABC posterior predictive.



(a) Predictive accuracy using energy score.

(b) kernelized Stein discrepancy.

FIG 7. ***Comparison between neural and linear stochastic parametrizations for the Lorenz96 model.*** *The posterior for the neural parametrization is sampled pSGLD algorithm more suited to high-dimensional spaces than the adSGLD used for the linear one. (a) Energy score between posterior predictive and each time-step of the original observation. This is repeated for 5 observations* $\mathbf{x_n}$ *(each using* $n = 1$ *here), and a t-distribution at each time-step is fitted to the energy score values. The solid line and shaded region respectively represent the mean and the 95% confidence interval of the fitted t-distribution. A lower energy score indicates better predictive performance. (b) KSD divided by the dimension of parameter space to assess the convergence of adSGLD for linear stochastic parametrization and pSGLD for neural stochastic parametrization.*

4.3.3. High dimensional neural stochastic parametrization for Lorenz96

The stochastic model considered in the previous section is a simplification of the original Lorenz96 model [53], which is a chaotic system including interacting slow and fast variables described by the following differential equations:

$$\begin{aligned} \frac{\mathrm{d}x_k}{\mathrm{d}t} &= -x_{k-1}\left(x_{k-2} - x_{k+1}\right) - x_k + F - \frac{hc}{b}\sum_{j=J(k-1)+1}^{kJ} y_j \\ \frac{\mathrm{d}y_j}{\mathrm{d}t} &= -\,\mathrm{cby}_{j+1}\left(y_{j+2} - y_{j-1}\right) - cy_j + \frac{hc}{b}X_{\mathrm{int}[(j-1)/J]+1}, \end{aligned} \tag{9}$$

where $k = 1, \ldots, K$, and $j = 1, \ldots, JK$, and cyclic boundary conditions are assumed, so that index $k = K + 1$ corresponds to $k = 1$ and similarly for j.

The stochastic model in Eq. (8) was derived by considering the part of the above ODE dealing with slow variables only and modelling the effect of the fast variables with the stochastic linear parametrization $g(y, t; \theta)$ [80]. To improve on this, we replace that with a high-dimensional parameterisation using a neural network:

$$g(x, t; \theta) = f(x; \theta) + \sigma_e \eta(t), \quad \eta(t) \sim \mathcal{N}(0, 1),$$

where $f(x; \theta)$ is a multi-layer perceptron with one hidden layer using a ReLU activation function. Altogether, this model has 111 parameters, on each of which we put an independent $\mathcal{N}(0, 10)$ prior.

To compare the linear and neural stochastic parametrizations, we simulate a time series from the full Lorenz96 model in equation (9) and consider this

as the observed data, by fixing $K = 8$, $J = 32$, $h = 1$, $b = 10$, $c = 10$ and $F = 10$. We then integrate the above equations with a 4th order Runge-Kutta integrator with $dt = 0.001$, starting from $x_k = y_j = 0$ for $k = 2, \dots, K$ and $j = 2, \dots JK$ and $x_1 = y_1 = 1$. We discard the first 2 time units and record the values of $\mathbf{x}$ every $\Delta t = 0.2$. This is done for a total of 21 timesteps. We repeat this process 5 times by perturbing the initial value with Gaussian noise; in this way, we generate 5 observations which slightly differ for the initial conditions (there is no other source of randomness as Eq. (9) is deterministic).

For the linearly parametrized Lorenz96 model, we follow the same setup as in Sec. 4.3.2 and use adSGLD to sample from the energy score posterior. In contrast, we opt to use pSGLD (Sec. 3.2) for the 111-dimensional neural Lorenz96 model. For both cases, we use $m = 500$ and 20000 MCMC steps. In Figure 7 we compare the inferred Scoring rule posterior via their predictive performance and convergence using KSD divided by the number of parameters (as the KSD grows linearly with the number of parameters). From this example, it is evident how SG-MCMC (more specifically pSGLD) enables sampling over a very high-dimensional parameter space very efficiently, which allows us to leverage a more expressive model to improve the representation of the observed data.

5. Related approaches

Scoring rules have been previously used to generalize Bayesian inference: [34] considered an update similar to ours, but fixed $w = 1$ and adjusted the parameter value (similarly to what was done in 67 and 74) so that the posterior has the same asymptotic covariance matrix as the frequentist minimum scoring rule estimator. Instead, [52] considered a time-series setting in which the task is to learn about the parameter value which yields the best prediction, given the previous observations. Finally, [41] motivated Bayesian inference using general divergences (beyond the KL one which underpins standard Bayesian inference) in an M-open setup, and discussed posteriors which employ estimators of the divergences from observed data; some of these estimators can be written using scoring rules. However, none of the above works considered explicitly the LFI setup.

A parallel work [56] investigates the generalized posterior obtained by using a kernelized Stein Discrepancy [14, 51]. This posterior is shown to satisfy robustness and consistency properties and is computationally convenient for doubly-intractable models (i.e., for which the likelihood is available, but only up to the normalizing constant). In contrast, our work focuses on models that do not have an explicit likelihood.

As mentioned before, previous LFI methods such as MMD-Bayes [13] and BSL [71] fall under our SR posterior framework. So do the semi-parametric BSL [1] and the ratio-estimation methods [77]; we discuss these methods in Appendices D.4 and D.5.

Interestingly, [21] introduced a new LFI method which, similar to ours, enjoys outlier robustness and posterior consistency; however, their method is derived

from the Bayesian non-parametric learning framework of [55, 28] rather than the generalized Bayesian posterior of [10].

Finally, [25] also uses Stochastic Gradient MCMC for sampling from a generalized posterior; however, instead of a reparametrization trick, the unbiased gradient estimate is obtained through a specific property of the system they consider (a quantum computer).

6. Conclusion

In this work, we introduced a generalized Bayesian posterior for likelihood-free inference relying on scoring rules which can easily be estimated with samples from the simulator model. This *scoring rule posterior* generalizes previous approaches [71, 13]. While pseudo-marginal MCMC enables approximate sampling of the posterior for simple cases, it mixes poorly for concentrated targets, even employing advanced schemes [68]; hence, we adapted Stochastic Gradient MCMC methods to our framework, by exploiting automatic differentiation to compute gradients for the simulator model. These new sampling schemes allow to sample the scoring rule posterior for high-dimensional parameter spaces. Our comparison with the popular Approximate Bayesian Computation and Bayesian Synthetic Likelihood showed how the scoring rule posterior enables more informative parameter inference, scaling to a higher number of samples and parameters. Moreover, it does so by using the raw data, while traditional LFI methods require determining suitable summary statistics.

We remark once again how the scoring rule posterior does *not* aim to approximate the standard Bayesian posterior, as most LFI methods do: it instead learns about the parameter value minimizing the expected scoring rule.

Appendix A: Proof of Theorem 3.1

We recall here for simplicity the useful definitions. We consider the SR posterior:

$$\pi_S(\theta|\mathbf{y_n}) \propto \pi(\theta) \underbrace{\exp\left\{-w \sum_{i=1}^{n} S(P_\theta, y_i)\right\}}_{p_S(\mathbf{y_n}|\theta)}.$$

Further, we recall the form of the target of the pseudo-marginal MCMC:

$$\pi_{\hat{S}}^{(m)}(\theta|\mathbf{y_n}) \propto \pi(\theta) p_{\hat{S}}^{(m)}(\mathbf{y_n}|\theta),$$

where:

$$\begin{aligned} p_{\hat{S}}^{(m)}(\mathbf{y_n}|\theta) &= \mathbb{E}\left[\exp\left\{-w \sum_{i=1}^{n} \hat{S}(\mathbf{X}_{\mathbf{m}}^{(\theta)}, y_i)\right\}\right] \\ &= \int \exp\left\{-w \sum_{i=1}^{n} \hat{S}(\mathbf{x}_{\mathbf{m}}^{(\theta)}, y_i)\right\} \prod_{j=1}^{m} p(x_j^{(\theta)}|\theta) dx_1 dx_2 \cdots dx_m. \end{aligned}$$

In order to prove Theorem 3.1, we extend the proof for the analogous result for Bayesian inference with an auxiliary likelihood [24]. Our setup is slightly more general as we do not constrain the update to be defined in terms of a likelihood; notice that the original setup in [24] is recovered when we consider S being the negative log likelihood, for some auxiliary likelihood.

We begin by stating a useful property:

Lemma A.1 (Theorem 3.5 in [8])**.** *If X_n is a sequence of uniformly integrable random variables and X_n converges in distribution to X, then X is integrable and $\mathbb{E}[X_n] \to \mathbb{E}[X]$ as $n \to \infty$.*

Remark 3 (Remark 1 in [24]). A simple sufficient condition for uniform integrability is that for some $\delta > 0$:

$$\sup_n \mathbb{E}[|X_n|^{1+\delta}] < \infty.$$

The result in the main text is the combination of the following two Theorems, which respectively generalize Results 1 and 2 in [24]:

Theorem A.2 (Generalizes Result 1 in [24])**.**
Assume that $p^{(m)}_{\hat{S}}(\mathbf{y_n}|\theta) \to p_S(\mathbf{y_n}|\theta)$ as $m \to \infty$ for all θ with positive prior support; further, assume $\inf_m \int_\Theta p^{(m)}_{\hat{S}}(\mathbf{y_n}|\theta)\pi(\theta)d\theta > 0$ and $\sup_m p^{(m)}_{\hat{S}}(\mathbf{y_n}|\theta) \leq g(\theta)$ where $\int_\Theta g(\theta)\pi(\theta)d\theta < \infty$. Then

$$\lim_{m\to\infty} \pi^{(m)}_{\hat{S}}(\theta|\mathbf{y_n}) = \pi_S(\theta|\mathbf{y_n}).$$

Furthermore, if $f : \Theta \to \mathbb{R}$ is a continuous function satisfying $\sup_m \int_\Theta |f(\theta)|^{1+\delta}\pi^{(m)}_S(\theta|\mathbf{y_n})d\theta < \infty$ for some $\delta > 0$ then

$$\lim_{m\to\infty} \int_\Theta f(\theta)\pi^{(m)}_{\hat{S}}(\theta|\mathbf{y_n})d\theta = \int_\Theta f(\theta)\pi_S(\theta|\mathbf{y_n})d\theta.$$

Proof. First, notice that, as $\inf_m \int_\Theta p^{(m)}_{\hat{S}}(\mathbf{y_n}|\theta)\pi(\theta)d\theta > 0$ and $\sup_m p^{(m)}_{\hat{S}}(\mathbf{y_n}|\theta) \leq g(\theta)$, the denominator of

$$\pi^{(m)}_{\hat{S}}(\theta|\mathbf{y_n}) = \frac{p^{(m)}_{\hat{S}}(\mathbf{y_n}|\theta)\pi(\theta)}{\int_\Theta p^{(m)}_{\hat{S}}(\mathbf{y_n}|\theta)\pi(\theta)d\theta}$$

is positive and converges by the dominated convergence theorem to $\int_\Theta \pi(\theta)p_S(\mathbf{y_n}|\theta)d\theta$. This, combines with the fact that the numerator converges pointwise, proves the first part.

For the second part, if for each $m \in \mathbb{N}, \theta_m$ is distributed according to $\pi^{(m)}_{\hat{S}}(\cdot|\mathbf{y_n})$ and θ is distributed according to $\pi_S(\cdot|\mathbf{y_n})$ then θ_m converges to θ in distribution as $m \to \infty$ by Scheffé's lemma [75]. Since f is continuous, $f(\theta_m)$ converges in distribution to $f(\theta)$ as $n \to \infty$ by the continuous mapping theorem and we conclude by application of Remark 3 and Lemma A.1. □

The following gives a convenient way to ensure $p^{(m)}_{\hat{S}}(\mathbf{y_n}|\theta) \to p_S(\mathbf{y_n}|\theta)$:

Theorem A.3 (Generalizes Result 2 in [24])**.** *Assume that* $\exp\{-w\sum_{i=1}^n \hat{S}(\mathbf{X}^{(\theta)}_{\mathbf{m}}, y_i)\}$ *converges in probability to* $p_S(\mathbf{y_n}|\theta)$ *as* $m \to \infty$*. If*

$$\sup_m \mathbb{E}\left[\exp\{-w(1+\delta)\sum_{i=1}^n \hat{S}(\mathbf{X}^{(\theta)}_{\mathbf{m}}, y_i)\}\right] < \infty$$

for some $\delta > 0$ *then* $p^{(m)}_{\hat{S}}(\mathbf{y_n}|\theta) \to p_S(\mathbf{y_n}|\theta)$ *as* $m \to \infty$*.*

Proof. First, notice that

$$\sup_m \mathbb{E}\left[\left|\exp\{-w\sum_{i=1}^n \hat{S}(\mathbf{X}^{(\theta)}_{\mathbf{m}}, y_i)\}\right|^{1+\delta}\right] = \sup_m \mathbb{E}\left[\exp\{-w(1+\delta)\sum_{i=1}^n \hat{S}(\mathbf{X}^{(\theta)}_{\mathbf{m}}, y_i)\}\right] < \infty.$$

The proof then follows by applying Remark 3 and Lemma A.1. □

We are finally ready to prove Theorem 3.1:

Proof of Theorem 3.1. First, notice how the convergence in probability of $\hat{S}(\mathbf{X}^{(\theta)}_{\mathbf{m}}, y_i)$ to $S(P_\theta, y_i)$ (assumption 1 in Theorem 3.1) and the continuity of the exponential function imply convergence in probability of $\exp\{-w\sum_i \hat{S}(\mathbf{X}^{(\theta)}_{\mathbf{m}}, y_i)\}$ to $p_S(\mathbf{y_n}|\theta)$. That, together with assumption 2 in Theorem 3.1, satisfy the requirements of Theorem A.3. With the latter and assumption 3 in Theorem 3.1, Theorem A.2 holds, which yields the result. □

A.1. Corollary when the SR estimator is lower bounded by the SR of the empirical distribution

This is a corollary of Theorem 3.1.

Corollary 1. Assume the following:

1. $\hat{S}(\mathbf{X}^{(\theta)}_{\mathbf{m}}, y_i)$ converges in probability to $S(P_\theta, y_i)$ as $m \to \infty$ for all $i = 1, \ldots, n$.
2. $\hat{S}(\mathbf{x}^{(\theta)}_{\mathbf{m}}, y_i) \geq S(\hat{P}_\theta, y_i)$, where $\hat{P}_\theta$ is the empirical distribution determined by $\mathbf{x}^{(\theta)}_{\mathbf{m}}$, and S is a proper scoring rule on the space of empirical distributions.
3. $\inf_m \int_\Theta p^{(m)}_{\hat{S}}(\mathbf{y_n}|\theta)\pi(\theta)d\theta > 0$ and $\sup_m p^{(m)}_{\hat{S}}(\mathbf{y_n}|\theta) \leq g(\theta)$ where $\int_\Theta g(\theta)\pi(\theta)d\theta < \infty$.

Then,

$$\lim_{m\to\infty} \pi^{(m)}_{\hat{S}}(\theta|\mathbf{y_n}) = \pi_S(\theta|\mathbf{y_n}).$$

Proof. We simply need to verify that Assumption 2 of the corollary verifies Assumption 2 of Theorem 3.1 (as the other two assumptions are preserved).

Let us denote by $\hat{Q}$ the empirical distribution obtained by $y_1, \dots, y_n$. As S is a proper scoring rule on the space of empirical distributions, it holds that

$$S(\hat{Q}, \hat{Q}) = \frac{1}{n}\sum_{i=1}^{n} S(\hat{Q}, y_i) \leq S(\hat{P}_\theta, \hat{Q}) = \frac{1}{n}\sum_{i=1}^{n} S(\hat{P}_\theta, y_i).$$

Combining this with the inequality in Assumption 2 of the corollary leads to

$$\frac{1}{n}\sum_{i=1}^{n} \hat{S}(\mathbf{x}_\mathbf{m}^{(\theta)}, y_i) \geq S(\hat{Q}, \hat{Q}),$$

which is a lower bound for fixed value of $\mathbf{y_n}$; in particular, this implies that

$$\exp\{-(1+\delta)w \sum_{i=1}^{n} \hat{S}(\mathbf{X}_\mathbf{m}^{(\theta)}, y_i)\} \leq \exp\{-(1+\delta)wn \cdot S(\hat{Q}, \hat{Q})\},$$

which implies Assumption 2. □

Notice that Assumption 2 above holds in the special case where the SR estimator corresponds to the SR of the empirical distribution.

Appendix B: Changing data coordinates

We give here some more details on the behavior of the SR posterior when the coordinate system used to represent the data is changed, as mentioned in Remark 2.

Frequentist estimator First, we investigate whether the minimum scoring rule estimator (for a strictly proper scoring rule) is affected by a transformation of the data. Specifically, considering a strictly proper S, we are interested in whether $\theta_Y^\star = \arg\min_{\theta\in\Theta} S(P_\theta^Y, Q_Y) = \arg\min_{\theta\in\Theta} D(P_\theta^Y, Q_Y)$ is the same as $\theta_Z^\star = \arg\min_{\theta\in\Theta} S(P_\theta^Z, Q_Z) = \arg\min_{\theta\in\Theta} D(P_\theta^Z, Q_Z)$, where $Z = f(Y) \implies Y \sim Q_Y \iff Z \sim Q_Z$ and $Y \sim P_\theta^Y \iff Z \sim P_\theta^Z$. If the model is well specified, $P_{\theta_Y^\star}^Y = Q_Y, P_{\theta_Z^\star}^Z = Q_Z \implies \theta_Y^\star = \theta_Z^\star$. If the model is misspecified, for a generic SR the minimizer of the expected SR may change according to the parametrization. We remark how this is not a drawback of the frequentist minimum SR estimator but rather a feature, as such estimator is the parameter value corresponding to the model minimizing the chosen expected scoring rule from the data generating process *in that coordinate system*, and is therefore completely reasonable for it to change when the coordinate system is modified.

Notice that a sufficient condition for $\theta_Y^\star = \theta_Z^\star$ is $S(P_\theta^Y, y) = a \cdot S(P_\theta^Z, z) + b$ for $a > 0, b \in \mathbb{R}$. This condition is verified when S is chosen to be the log-score, as in fact:

$$S(P_\theta^Z, f(y)) = -\ln p_Z(f(y)|\theta) = S(P_\theta^Z, y) + \ln|J_f(y)|,$$

where we assumed f to be a one-to-one function and we applied the change of variable formula to the density p_Z.

Generalized Bayesian posterior For a single observation, let π_S^Y denote the SR posterior conditioned on values of Y, while π_S^Z denote instead the posterior conditioned on values of $Z = f(Y)$ for some one-to-one function f; in general, $\pi_S^Y(\theta|y) \neq \pi_S^Z(\theta|f(y))$. By denoting as w_Z (respectively w_Y) and P_θ^Z (respectively P_θ^Y) the weight and model distributions appearing in π_S^Z (resp. π_S^Y), the equality would in fact require $w_Z S(P_\theta^Z, f(y)) = w_Y S(P_\theta^Y, y) + C \; \forall \, \theta, y$ for some choice of w_Z, w_Y and for all transformations f, where C is a constant in θ. Notice that this is satisfied for the standard Bayesian posterior (i.e., with the log-score) with $w_Z = w_Y = 1$. Instead, for other scoring rules the above condition cannot be satisfied in general for any choice of w_Z, w_Y. For instance, consider the kernel SR:

$$S(P_\theta^Z, f(y)) = \mathbb{E}[k(Z, \tilde{Z})] - \mathbb{E}[k(Z, f(y))] = \mathbb{E}[k(f(Y), f(\tilde{Y}))] - \mathbb{E}[k(f(Y), f(y))];$$

for general kernels and functions f, the above is different from $S(P_\theta^Y, y) = \mathbb{E}[k(Y, \tilde{Y})] - \mathbb{E}[k(Y, f(x))]$ up to a constant, unless the kernel is redefined as well. Therefore, the posterior shape depends on the chosen data coordinates. Considering the expression for the kernel SR, it is clear that is a consequence of the fact that the likelihood principle is not satisfied (as the kernel SR does not only depend on the likelihood value at the observation). Similar argument holds for the energy score posterior as well.

We also remark that this is also the case for BSL [71], as in that case the model is assumed to be multivariate normal, and changing the data coordinates impacts their normality (in fact it is common practice in BSL to look for transformations of data which yield distribution as close as possible to a normal one).

The theoretical semiBSL posterior [1], instead, is invariant with respect to one-to-one transformation applied independently to each data coordinate, which do not affect the copula structure. Notice however that different data coordinate systems may yield better empirical estimates of the marginal KDEs from model simulations.

Appendix C: Checking convergence of MCMC with the kernelized Stein discrepancy

As SG-MCMC algorithms in general exhibit an asymptotic bias, we require a convergence test which accounts for this bias in the stationary distribution. We thus utilise the method of Kernelized Stein Discrepancy (KSD) proposed in [36], which is especially applicable in the case of stochastic gradient MCMC as it depends on the target distribution only through its gradient.

Given the samples of our parameter $\{\theta_1, ..., \theta_n\}$ where $\theta_i \in \mathbb{R}^d$, we denote the empirical distribution described by these samples as $\tilde{\pi}$, and our target distribution as π. We consider the Integral Probability Metric (IPM) defined over a class of test function $\mathcal{H}$,

$$d_{\mathcal{H}}(\tilde{\pi}, \pi) := \sup_{h \in \mathcal{H}} \left| \mathbb{E}_{\tilde{\pi}}[h(\theta)] - \mathbb{E}_{\pi}[h(\theta)] \right|.$$

For IPMs such as the Wasserstein distance, we obtain a desirable property that $d_{\mathcal{H}}(\tilde{\pi}_K, \pi) \to 0$ implies $\tilde{\pi}_K \Rightarrow \pi$ (weak convergence of measures). However, since π is not available for integration, we instead utilise a class of IPMs called Stein Discrepancy, constructed such that the test functions give zero mean under π. We do this by defining a Stein operator, $\mathcal{T}$, which maps functions g: $\mathbb{R}^d \to \mathbb{R}^d$ from our Stein set, the domain $\mathcal{G}$. This is chosen such that $\mathbb{E}_\pi[(\mathcal{T}g)(Z)] = 0$ for all $g \in \mathcal{G}$. Then we can define the Stein discrepancy:

$$\begin{aligned}\mathcal{S}(\tilde{\pi}, \mathcal{T}, \mathcal{G}) := d_{\mathcal{TG}}(\tilde{\pi}, \pi) &= \sup_{g \in \mathcal{G}} |\mathbb{E}_{\tilde{\pi}}[(\mathcal{T}g)(X)] - \mathbb{E}_\pi[(\mathcal{T}g)(Z)]| \\ &= \sup_{g \in \mathcal{G}} |\mathbb{E}_{\tilde{\pi}}[(\mathcal{T}g)(X)]| \,.\end{aligned}$$

Thus, such a Stein operator and Stein set must be chosen to fulfil the Stein discrepancy condition and the desired convergence property. In [36], the Stein operator is proposed to be the Langevin Stein operator,

$$(\mathcal{T}_P g)(x) := \langle g(x), \nabla \log p(x) \rangle + \langle \nabla, g(x) \rangle$$

and the corresponding Stein set, which is defined using a Reproducing Kernel Hilbert space of function $\mathcal{K}_k$. We denote $\|\cdot\|_{\mathcal{K}_k}$ to be the induced norm from the inner product in $\mathcal{K}_k$, and $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the reproducing kernel of $\mathcal{K}_k$. This is the kernelized Stein set:

$$\mathcal{G}_{k,\|\cdot\|} := \left\{ g = (g_1, \ldots, g_d) \mid \|v\|^* \leq 1 \text{ for } v_j := \|g_j\|_{\mathcal{K}_k} \right\}$$

where $g = (g_1, \ldots, g_d)$ is a vector-valued function. This combination of the Langevin Stein operator and the kernelized Stein set is known as the kernelized Stein Discrepancy (KSD) $S\left(\mu, \mathcal{T}_P, \mathcal{G}_{k,\|\cdot\|}\right)$, for a probability measure μ. In [36], the KSD was proven to have a closed form solution for any $\|\cdot\|$, which of particular interest to us is when $S\left(\tilde{\pi}, \mathcal{T}_P, \mathcal{G}_{k,\|\cdot\|}\right)$,

$$S\left(\tilde{\pi}, \mathcal{T}_P, \mathcal{G}_{k,\|\cdot\|}\right) := \sum_{j=1}^{d} \sqrt{\sum_{i,i'=1}^{n} \frac{k_j^0(\theta_i, \theta_{i'})}{n^2}}$$

where the Stein kernel for $j \in \{1, \ldots, d\}$ is given by

$$\begin{aligned}k_j^0(\theta, \theta') = &\left(\nabla_{\theta^{(j)}} U(\theta) \nabla_{\theta^{(j)}} U(\theta')\right) k(\theta, \theta') + \nabla_{\theta^{(j)}} U(\theta) \nabla_{\theta'^{(j)}} k(\theta, \theta') \\ &+ \nabla_{\theta'^{(j)}} U(\theta') \nabla_{\theta^{(j)}} k(\theta, \theta') + \nabla_{\theta^{(j)}} \nabla_{\theta^{(j)}} k(\theta, \theta') \,,\end{aligned}$$

where $U(\theta)$ is such that $\pi(\theta) \propto e^{-U(\theta)}$. Note that [36] recommended the use of the inverse multi quadric kernel, $k(\theta, \theta') = \left(c^2 + \|\theta - \theta'\|_2^2\right)^\beta$ which gives desired convergence properties when $c > 0$ and $\beta \in (-1, 0)$.

In our specific case of the SR posterior, $U(\theta) = w \cdot \sum_{i=1}^{n} S(P_\theta, y_i)$. As for the energy and kernel scores we cannot exactly evaluate $\nabla_\theta U(\theta)$, we reaplce it with an unbiased estimate when computing the KSD.

Appendix D: More details on related techniques

D.1. Energy distance

The squared energy distance is a metric between probability distributions [72], and is defined by:

$$D_{\mathrm{E}}^{(\beta)}(P,Q) = 2 \cdot \mathbb{E}\left[\|X-Y\|_2^\beta\right] - \mathbb{E}\left[\|X-X'\|_2^\beta\right] - \mathbb{E}\left[\|Y-Y'\|_2^\beta\right],$$

for $X \perp\!\!\!\perp X' \sim P$ and $Y \perp\!\!\!\perp Y' \sim Q$.

The probabilistic forecasting literature [35] use a different convention of the energy score and distance, which amounts to multiplying our definitions by 1/2. We follow here the convention used in the statistical inference literature [72, 13, 60].

D.2. Maximum Mean Discrepancy (MMD)

We follow here Section 2.2 in [37]; all proofs of our statements can be found there. Let $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive-definite and symmetric kernel; notice that this implies $k(x,x) \geq 0$. Under these conditions, there exists a unique Reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ of real functions on $\mathcal{X}$ associated to k.

Now, let's define the Maximum Mean Discrepancy (MMD).

Definition D.1. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \to \mathbb{R}$; we define the MMD relative to $\mathcal{F}$ as:

$$\mathrm{MMD}_{\mathcal{F}}(P,Q) = \sup_{f\in\mathcal{F}} \left[\mathbb{E}_{X\sim P} f(X) - \mathbb{E}_{Y\sim Q} f(Y)\right].$$

We will show here how choosing $\mathcal{F}$ to be the unit ball in an RKHS $\mathcal{H}_k$ turns out to be computationally convenient, as it allows us to avoid computing the supremum explicitly. First, let us define the mean embedding of the distribution P in $\mathcal{H}_k$:

Lemma D.2 (Lemma 3 in [37])**.** *If $k(\cdot,\cdot)$ is measurable and $\mathbb{E}_{X\sim P}\sqrt{k(X,X)} < \infty$, then the mean embedding of the distribution P in $\mathcal{H}_k$ is:*

$$\mu_P = \mathbb{E}_{X\sim P}\left[k(X,\cdot)\right] \in \mathcal{H}_k.$$

Using this fact, the following Lemma shows that the MMD relative to $\mathcal{H}_k$ can be expressed as the distance in $\mathcal{H}_k$ between the mean embeddings:

Lemma D.3 (Lemma 4 in [37])**.** *Assume the conditions in Lemma D.2 are satisfied, and let $\mathcal{F}$ be the unit ball in $\mathcal{H}_k$; then:*

$$\mathrm{MMD}_{\mathcal{F}}^2(P,Q) = ||\mu_P - \mu_Q||_{\mathcal{H}}^2.$$

In general, the MMD is a *pseudo-metric* for probability distributions (i.e., it is symmetric, satisfies the triangle inequality and $\mathrm{MMD}_{\mathcal{F}}(P,P) = 0$, 12). For the probability measures on a compact metric space $\mathcal{X}$, the next Lemma states the conditions under which the MMD is a *metric*, which additionally ensures that $\mathrm{MMD}_{\mathcal{F}}(P,Q) = 0 \implies P = Q$. Specifically, this holds when the kernel is universal, which requires that $k(\cdot,\cdot)$ is continuous, and $\mathcal{H}_k$ being dense in $C(\mathcal{X})$ with respect to the L_∞ norm (these conditions are satisfied by the Gaussian and Laplace kernel).

Lemma D.4 (Theorem 5 in [37])**.** *Let $\mathcal{F}$ be the unit ball in $\mathcal{H}_k$, where $\mathcal{H}_k$ is defined on a compact metric space $\mathcal{X}$ and has associated continuous kernel $k(\cdot,\cdot)$. Then:*

$$\mathrm{MMD}_{\mathcal{F}}(P,Q) = 0 \iff P = Q.$$

This result can be generalized to more general spaces $\mathcal{X}$, by considering the notion of characteristics kernel, for which the mean map is injective; it can be shown that the Laplace and Gaussian kernels are characteristics [37], so that MMD for those two kernels is a metric for distributions on $\mathbb{R}^d$.

Additionally, the form of MMD for a unit-ball in an RKHS allows easy estimation, as shown next:

Lemma D.5 (Lemma 6 in [37])**.** *Assume that the form for MMD given in Lemma D.3 holds; say $X \perp\!\!\!\perp X' \sim P$, $Y \perp\!\!\!\perp Y' \sim Q$, and let $\mathcal{F}$ be the unit ball in $\mathcal{H}_k$. Then, you can write:*

$$\mathrm{MMD}^2_{\mathcal{F}}(P,Q) = \mathbb{E}[k(X,X')] + \mathbb{E}[k(Y,Y')] - 2\mathbb{E}[k(X,Y)].$$

D.2.1. Equivalence between MMD-Bayes posterior and π_{S_k}

[13] considered the following posterior, termed MMD-Bayes:

$$\pi_{\mathrm{MMD}}(\theta|\mathbf{y_n}) \propto \pi(\theta)\exp\left\{-\beta \cdot D_k\left(P_\theta, \hat{P}_n\right)\right\}$$

where $\beta > 0$ is a temperature parameter and $D_k\left(P_\theta, \hat{P}_n\right)$ denotes the squared MMD between the empirical measure of the observations $\hat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$ and the model distribution P_θ.

From the properties of MMD (see Appendix D.2), notice that:

$$\begin{aligned} D_k\left(P_\theta, \hat{P}_n\right) &= \mathbb{E}_{X,X'\sim P_\theta} k(X,X') + \frac{1}{n^2}\sum_{i,j=1}^n k(y_i,y_j) - \frac{2}{n}\sum_{i=1}^n \mathbb{E}_{X\sim P_\theta} k(X,y_i) \\ &= \frac{1}{n}\left(n\cdot \mathbb{E}_{X,X'\sim P_\theta} k(X,X') - 2\sum_{i=1}^n \mathbb{E}_{X\sim P_\theta} k(X,y_i)\right) + \frac{1}{n^2}\sum_{i,j=1}^n k(y_i,y_j) \\ &= \frac{1}{n}\left(\sum_{i=1}^n S_k(P_\theta, y_i)\right) + \frac{1}{n^2}\sum_{i,j=1}^n k(y_i,y_j), \end{aligned}$$

where we used the expression of the SR scoring rule S_k, and where the second term is independent on θ. Therefore, the MMD-Bayes posterior is equivalent to the SR posterior with kernel scoring rule S_k, by identifying $w = \beta/n$.

D.3. The Dawid–Sebastiani score

As mentioned in Sec. 4.3.1, the BSL posterior can be seen as a scoring rule posterior with $w = 1$ considering the Dawid–Sebastiani (DS) score, which is defined as:

$$S_{\mathrm{DS}}(P, y) = \ln |\Sigma_P| + (y - \mu_P)^T \Sigma_P^{-1} (y - \mu_P),$$

where μ_P and Σ_P are the mean vector and covariance matrix of P. The DS score is the negative log-likelihood of a multivariate normal distribution with mean μ_P and covariance matrix Σ_P, up to some constants. Therefore, it is equivalent to the log score when P is a multivariate normal distribution. For a set of distributions $\mathcal{P}(\mathcal{X})$ with well-defined second moments, this SR is proper but not strictly so: several distributions of that class may yield the same score, as long as the two first moments match [35]. It is strictly proper if distributions in $\mathcal{P}(\mathcal{X})$ are determined by their first two moments, as it is the case for the normal distribution.

D.4. Semi-parametric synthetic likelihood

We review here the semiBSL approach [1].

Copula theory First, recall that a copula is a multivariate Cumulative Density Function (CDF) such that the marginal distribution for each variable is uniform on the interval $[0, 1]$. Consider now a multivariate random variable $X = (X^1, \ldots, X^d)$, for which the marginal CDFs are denoted by $F_j(x) = \mathbb{P}(X^j < x)$; then, the multivariate random variable built as:

$$(U^1, U^2, \ldots, U^d) = (F_1(X^1), F_2(X^2), \ldots, F_d(X^d))$$

has uniform marginals on $[0, 1]$.

Sklar's theorem exploits copulas to decompose the density h of X[6]; specifically, it states that the following decomposition is valid:

$$h(x^1, \ldots, x^d) = c(F_1(x^1), \ldots, F_d(x^d)) f_1(x^1) \cdots f_d(x^d),$$

where f_j is the marginal density of the j-th coordinate, and c is the density of the copula.

We now review definition and properties of the Gaussian copula, which is defined by a correlation matrix $R \in [-1, 1]^{d \times d}$, and has cumulative density function:

$$C_R(u) = \Phi_R(\Phi^{-1}(u^1), \ldots, \Phi^{-1}(u^d)),$$

[6]Provided that the density exists in the first place; a more general version of Sklar's theorem is concerned with general random variables, but we restrict here to the case where densities are available.

where Φ^{-1} is the inverse cdf (quantile function) of a standard normal, and Φ_R is the cdf of a multivariate normal with covariance matrix R and 0 mean. If you define as U the random variable which is distributed according to C_R, it can easily be seen that R is the covariance matrix of the multivariate normal random variable $Z = \Phi^{-1}(U)$, where Φ^{-1} is applied element-wise. In fact:

$$P(Z \leq \eta) = P(U \leq \Phi(\eta)) = C_R(\Phi(\eta)) = \Phi_R(\eta),$$

where the inequalities are intended component-wise.

By defining as η a d-vector with components $\eta^k = \Phi^{-1}(u^k)$, the Gaussian copula density is:

$$c_R(u) = \frac{1}{\sqrt{|R|}} \exp\left\{-\frac{1}{2}\eta^\top \left(R^{-1} - \mathbf{I}_d\right)\eta\right\},$$

where $\mathbf{I}_d$ is a d-dimensional identity matrix, and $|\cdot|$ denotes the determinant.

Semiparametric Bayesian Synthetic Likelihood (semiBSL) SemiBSL assumes that the likelihood for the model has a Gaussian copula; therefore, the likelihood for a single observation y can be written as:

$$p_{\text{semiBSL}}(y|\theta) = c_{R_\theta}(F_{\theta,1}(y^1), \ldots, F_{\theta,d}(y^d)) \prod_{k=1}^{d} f_{\theta,k}\left(y^k\right),$$

where y^k is the k-th component of y, $f_{\theta,k}$ is the marginal density of the k-th component and $F_{\theta,k}$ is the CDF of the k-th component.

In order to obtain an estimate for it, we exploit simulations from P_θ to estimate R_θ, $f_{\theta,k}$ and $F_{\theta,k}$; this leads to:

$$\begin{aligned}\hat{p}_{\text{semiBSL}}(y|\theta) &= c_{\hat{R}_\theta}(\hat{F}_{\theta,1}(y^1), \ldots, \hat{F}_{\theta,d}(y^d)) \prod_{k=1}^{d} \hat{f}_{\theta,k}\left(y^k\right) \\ &= \frac{1}{\sqrt{|\hat{R}_\theta|}} \exp\left\{-\frac{1}{2}\hat{\eta}_y^\top \left(\hat{R}_\theta^{-1} - \mathbf{I}_d\right)\hat{\eta}_y\right\} \prod_{k=1}^{d} \hat{f}_{\theta,k}\left(y^k\right),\end{aligned}$$

where $\hat{f}_{\theta,k}$ and $\hat{F}_{\theta,k}$ are estimates for $f_{\theta,k}$ and $F_{\theta,k}$, $\hat{\eta}_y = (\hat{\eta}_y^1, \ldots, \hat{\eta}_y^d)$, $\hat{\eta}_y^k = \Phi^{-1}(\hat{u}^k)$, $\hat{u}^k = \hat{F}_{\theta,k}(y^k)$. Moreover, $\hat{R}_\theta$ is an estimate of the correlation matrix.

We discuss now how the different quantities are estimated. First, a Kernel Density Estimate (KDE) is used for the marginals densities and cumulative density functions. Specifically, given samples $x_1, \ldots, x_m \sim P_\theta$, a KDE estimate for the k-th marginal density is:

$$\hat{f}_{\theta,k}(y^k) = \frac{1}{m}\sum_{j=1}^{m} K_h(y^k - x_j^k),$$

where K_h is a normalized kernel which is chosen to be Gaussian in the original implementation [1]. The CDF estimates are obtained by integrating the KDE density.

Next, for estimating the correlation matrix, [1] proposed to use a robust procedure based on the ranks (grc, Gaussian rank correlation, 11); specifically, given m simulations $x_1, \dots, x_m \sim P_\theta$, the estimate for the (k,l)-th entry of R_θ is given by:

$$\left[\hat{R}_\theta^{\text{grc}}\right]_{k,l} = \frac{\sum_{j=1}^m \Phi^{-1}\left(\frac{r\left(x_j^k\right)}{m+1}\right)\Phi^{-1}\left(\frac{r\left(x_j^l\right)}{m+1}\right)}{\sum_{j=1}^m \Phi^{-1}\left(\frac{j}{m+1}\right)^2},$$

where $r(\cdot) : \mathbb{R} \to \mathcal{A}$, where $\mathcal{A} = \{1, \dots, m\}$ is the rank function.

Copula scoring rule Finally, we write down the explicit expression of the copula scoring rule S_{Gc}, associated to the Gaussian copula. We show that this is a proper, but not strictly so, scoring rule for copula distributions. Specifically, let C be a distribution for a copula random variable, and let $u \in [0,1]^d$. We define:

$$S_{Gc}(C,u) = \frac{1}{2}\log|R_C| + \frac{1}{2}\left(\Phi^{-1}(u)\right)^T (R_C^{-1} - \mathbf{I}_d)\Phi^{-1}(u),$$

where Φ^{-1} is applied element-wise to u, and R_C is the correlation matrix associated to C in the following way: define the copula random variable $V \sim C$ and its transformation $\Phi^{-1}(V)$; then, $\Phi^{-1}(V)$ will have a multivariate normal distribution with mean 0 and covariance matrix R_C.

Similarly to the Dawid–Sebastiani score (see Appendix D.3), this scoring rule is proper but not strictly so as it only depends on the first 2 moments of the distribution of the random variable $\Phi^{-1}(V)$ (the first one being equal to 0). To show this, assume the copula random variable U has an exact distribution Q and consider the expected scoring rule:

$$\begin{aligned} S_{Gc}(C,Q) &= \mathbb{E}_{U\sim Q} S_{Gc}(C,U) \\ &= \frac{1}{2}\log|R_C| + E_{U\sim Q}\left[\left(\Phi^{-1}(U)\right)^T (R_C^{-1} - \mathbf{I}_d)\Phi^{-1}(U)\right]; \end{aligned}$$

now, notice that $\Phi^{-1}(U)$ is a multivariate normal distribution whose marginals are standard normals. Therefore, let us denote as R_Q the covariance matrix of $\Phi^{-1}(U)$, which is a correlation matrix. From the well-known form for the

expectation of a quadratic form[7], it follows that:

$$\begin{aligned} S_{Gc}(C,Q) &= \frac{1}{2}\log|R_C| + \frac{1}{2}\mathrm{Tr}\left[(R_C^{-1} - \mathbf{I}_d)\cdot R_Q\right] \\ &= \frac{1}{2}\log|R_C| + \frac{1}{2}\mathrm{Tr}\left[R_C^{-1}\cdot R_Q\right] - \frac{1}{2}\mathrm{Tr}\left[R_Q\right] \\ &= \underbrace{\frac{1}{2}\left\{\log\frac{|R_C|}{|R_Q|} - d + \mathrm{Tr}\left[R_C^{-1}\cdot R_Q\right]\right\}}_{D_{KL}(Z_Q||Z_C)} + \frac{1}{2}\log R_Q + \frac{d}{2} - \frac{1}{2}\mathrm{Tr}\left[R_Q\right], \end{aligned}$$

where $D_{KL}(Z_Q||Z_C)$ is the KL divergence between two multivariate normal distributions Z_Q and Z_C of dimension d, with mean 0 and covariance matrix R_Q and R_C respectively. Further, notice that the remaining factors do not depend on the distribution C. Therefore, $S_{Gc}(C,Q)$ is minimized whenever R_C is equal to R_Q; this happens when $C = Q$, but also for all other choices of C which share the associated covariance matrix with Q. This implies that the Gaussian copula score is a proper, but not strictly so, scoring rule for copula distributions.

D.5. Ratio estimation

The standard Bayes posterior can be written as $\pi(\theta|y) = \pi(\theta)\cdot r(y;\theta)$, with $r(y;\theta) = \frac{p(y|\theta)}{p(y)}$. The Ratio Estimation (RE) approach [77] builds an approximate posterior by estimating $\log r(y;\theta)$ with some function $\hat{h}^\theta(y)$ and considering $\pi_{\mathrm{re}}(\theta|y) \propto \pi(\theta)\exp(\hat{h}^\theta(y))$.

[77] run an MCMC where, for each proposed θ, m samples $\mathbf{x}_\mathbf{m}^{(\theta)}$ are generated from P_θ. These, together with a set of m reference samples $\mathbf{x}_\mathbf{m}^{(r)} = \{x_j^{(r)}\}_{j=1}^m$ from the marginal data distribution[8], are used to fit a logistic regression yielding $\hat{h}^\theta(y)$. Logistic regression is an optimization problem in which the best function of $\mathcal{X}$ in distinguishing between the two sets of samples is selected. If $m \to \infty$ and all scalar functions are considered, the optimum $h_\star^\theta$ is equal to $\log r(y;\theta)$. For finite data, however, the corresponding optimum $\hat{h}_m^\theta$ is only an approximation of the ratio (as discussed in Appendix D.5). RE is therefore a specific case of our SR posterior framework with $w = 1$ and:

$$\hat{S}_{\mathrm{RE}}(\mathbf{x}_\mathbf{m}^{(\theta)}, \mathbf{x}_\mathbf{m}^{(r)}, y) = -\hat{h}_m^\theta(y)$$

which, differently from the other SR estimators considered previously, also depends on the reference samples. Due to what we discussed above, $\hat{S}_{\mathrm{RE}}$ converges in probability to the log-score (up to a constant term in θ) for $m \to \infty$.

[7] $\mathbb{E}\left[X^T\Lambda X\right] = \mathrm{tr}\left[\Lambda\Sigma\right] + \mu^T\Lambda\mu$, for a symmetric matrix Λ, and where μ and Σ are the mean and covariance matrix of X (which in general does not need to be normal, but only needs to have well defined second moments).

[8] Which are obtained by drawing $\theta_j \sim p(\theta)$, $x_j \sim p(\cdot|\theta_j)$, and discarding θ_j. In general, the number of reference samples and samples from the model can be different, see Appendix D.5; we make this choice here for the sake of simplicity.

The above argument relies on optimizing over all functions in logistic regression; in practice, the optimization is restricted to a set of functions $\mathcal{H}$ (for instance, a linear combination of predictors). In this case, the infinite data optimum $h^{\theta}_{\mathcal{H}^\star}(y)$ does not correspond to $\log r(y;\theta)$ (see Appendix D.5), but to the best possible approximation in $\mathcal{H}$ in some sense. Therefore, Ratio Estimation with a restricted set of functions $\mathcal{H}$ cannot be written exactly under our SR posterior framework. However, very flexible function classes (as for instance neural networks) can produce reasonable approximations to the log score for large values of m.

Appendix E: Tuning the bandwidth of the Gaussian kernel

Consider the Gaussian kernel:

$$k(x,y) = \exp\left(-\frac{\|x-y\|_2^2}{2\gamma^2}\right);$$

inspired by [65], we fix the bandwidth γ with the following procedure:

1. Simulate a value $\theta_j \sim \pi(\theta)$ and a set of samples $x_{jk} \sim P_{\theta_j}$, for $k = 1,\dots,m_\gamma$.
2. Estimate the median of $\{||x_{jk} - x_{jl}||_2\}_{kl}^{m_\gamma}$ and call it $\hat{\gamma}_j$.
3. Repeat points 1) and 2) for $j = 1,\dots,m_{\theta,\gamma}$.
4. Set the estimate for γ as the median of $\{\hat{\gamma}_j\}_{j=1}^{m_{\theta,\gamma}}$.

Empirically, we use $m_{\theta,\gamma} = 1000$ and we set m_γ to the corresponding value of m for the different models.

Appendix F: Further details on simulation studies reported in the main text

F.1. The SR posterior on the g-and-k model

We report here additional experimental details on the g-and-k model experiments.

F.1.1. The SR posterior on univariate g-and-k

SG-MCMC and PM-MCMC comparison We ran our inference with observations of $n = 10$. Both energy score posteriors for PM-MCMC and SG-MCMC was set to $w = 1$.

- For the SR posterior with SG-MCMC, we utilised the adSGLD algorithm, with the step-size ϵ tuned with the Multi-Armed Bandit algorithm [16] as discussed previously. The chain was initialized at a parameter value of 0. This resulted in $\epsilon = 3 \times 10^{-3}$.
- For the SR posterior with PM-MCMC, we utilised a proposal size of $\sigma = 1$.

Concentration study For our concentration study, we ran our inference with increasing observations of $n = 1, 10, 20, 50, 70, 100, 200$. Generally, we ran the chain with the Multi-Armed Bandit algorithm [16] as discussed previously. The chains were started from an initial optimization step of 250 iterations ran with the Adam optimizer [44]. In Table 1, we report the final step-size determined by the Multi-Armed Bandit algorithm for different values of n. We detail below the settings for the different SR posteriors.

- The energy score posteriors were set to $w = 1$.
- For the kernel score posteriors, we set w using our heuristic procedure discussed in Sec. 4.2 with the energy score posterior as a reference, resulting in $w = 28.1$. The Gaussian kernel bandwidth γ, was tuned using the procedure detailed in Appendix E, resulting in $\gamma = 5.47$.

TABLE 1
Step-sizes for the two SR posteriors in the univariate g-and-k model, determined with the Multi-Armed Bandit algorithm of [16].

Observations	$n = 1$	$n = 10$	$n = 20$	$n = 50$	$n = 70$	$n = 100$	$n = 200$	$n = 400$
Energy score	3×10^{-2}	3×10^{-2}	3×10^{-3}	3×10^{-4}	1×10^{-3}	1×10^{-4}	1×10^{-4}	3×10^{-6}
Kernel score	1×10^{-1}	3×10^{-2}	1×10^{-2}	1×10^{-3}	1×10^{-3}	1×10^{-4}	3×10^{-5}	3×10^{-5}

F.1.2. The SR posterior on multivariate g-and-k

Similar to the univariate model, we ran our inference with increasing observations of $n = 1, 10, 20, 50, 70, 100, 200, 400$ and with the Multi-Armed Bandit algorithm [16] as discussed previously. The chains were started from an initial optimization step of 250 iterations ran with the Adam optimizer [44]. In Table 2 and Table 3, we report the final step-size determined by the Multi-Armed Bandit algorithm for different values of n for the well-specified case and the misspecified case respectively.

We detail below the settings for the different SR posteriors and for the BSL posterior.

Well-specified case

- The energy score posteriors were set to $w = 1$.
- For the kernel score posteriors, we set w using our heuristic procedure discussed earlier with the energy score posterior as a reference, resulting in $w = 191$. The Gaussian kernel bandwidth γ, was tuned using the procedure detailed in E, resulting in $\gamma = 45$.
- For the BSL posteriors, we set $\sigma = 1$. However, the chain was unable to converge for any $n > 10$, and so we ran the BSL posterior with an additional $n = 5$ observations.

TABLE 2
Step-sizes for the two SR posteriors in the multivariate g-and-k model with well-specified observations, determined with the Multi-Armed Bandit algorithm of [16].

Observations	$n = 1$	$n = 10$	$n = 20$	$n = 50$	$n = 70$	$n = 100$	$n = 200$	$n = 400$
Energy score	1×10^{-1}	3×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-4}	3×10^{-5}	1×10^{-5}	3×10^{-6}
Kernel score	1×10^{-1}	3×10^{-4}	3×10^{-4}	1×10^{-4}	1×10^{-5}	3×10^{-6}	1×10^{-5}	1×10^{-6}

Misspecified case Due to the misspecified model, for certain values of n, the SG-MCMC algorithm resulted in proposal values that were outside our specified parameter range. For these cases, we manually tuned the step-size such that the SG-MCMC algorithm ran successfully. These cases are indicated in Table 3 with an asterisk $(*)$.

- The energy score posteriors were set to $w = 1$.
- For the kernel score posteriors, in order to have coherent results with respect to the well-specified case, we use here the values determined in the well-specified case. ($w = 191$, $\gamma = 45$)
- For the BSL posteriors, we set $\sigma = 1$. However, the chain was unable to converge for any $n > 5$, and so we ran the BSL posterior with an additional $n = 5$ observations.

TABLE 3
Step-sizes for the two SR posteriors in the multivariate g-and-k model with misspecified observations, determined with the Multi-Armed Bandit algorithm of [16].

Observations	$n = 1$	$n = 10$	$n = 20$	$n = 50$	$n = 70$	$n = 100$	$n = 200$	$n = 400$
Energy score	1×10^{-1}	1×10^{-2}	3×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-4}	1×10^{-4}	3×10^{-5}
Kernel score	5×10^{-2} (*)	1×10^{-2}	3×10^{-3}	3×10^{-3}	3×10^{-3}	3×10^{-3}	1×10^{-4} (*)	6×10^{-5} (*)

F.2. The Lorenz96 model

In both comparisons, we utilize the energy score posterior with $w = 1$, except for the case where the SMC-ABC algorithm is used.

Comparison with ABC We ran the inference using the adSGLD algorithm, and the SMC-ABC algorithm, both with observations of $n = 10$. For the energy score posterior, a step size of $\epsilon = 3 \times 10^{-2}$ was set, and the chain was initialised at a parameter value of 0.

High dimensional neural stochastic parametrization We ran the inference using both the adSGLD algorithm with the linear stochastic parametrization and the pSGLD algorithm with the high dimensional neural parametrization, both with observations of $n = 1$ which was first standardised. For both cases, chains were started from an initial optimization step of 250 iterations ran with the Adam optimizer [44]. For the adSGLD algorithm, a step size of $\epsilon = 1 \times 10^{-4}$ was set, while for the pSGLD algorithm this was set to $\epsilon = 1 \times 10^{-7}$.

Appendix G: Results with pseudo-marginal MCMC on g-and-k model

We report here some parallel results to those in the main text of the paper obtained with pseudo-marginal (PM) MCMC. To obtain these results, we use the correlated pseudo-marginal MCMC [17, 20, 68] mentioned in Sec. 3.1 with independent normal proposals on each component of the parameter space; we indicate by σ the standard deviation of the normal proposal distribution, which we report below. In all cases, whenever the parameter space is bounded, we run PM-MCMC on a transformed unbounded space obtained via a logistic transformation. Therefore, the proposal sizes refer to that unbounded space.

Besides our SR posteriors, we consider here the BSL and the semi-parametric BSL (Appendix D.4). When performing these studies, we aimed at comparing the performance of our SR posteriors with BSL. BSL was run both with and without summaries (see Section 4.3.1), while semi-parametric BSL was only run without summaries. Both semi-parametric BSL and BSL with summaries were only run for the multivariate case (notice that the former is only well-defined for multivariate models). However, in this appendix, we only discuss results with summary-free BSL. Hence, we set the value of w for the energy and kernel score posteriors with the strategy discussed in Sec. 4.2 using BSL as a reference.

G.1. Well-specified setup

For both univariate and multivariate case, we consider synthetic observations generated from parameter values $A^\star = 3$, $B^\star = 1.5$, $g^\star = 0.5$, $k^\star = 1.5$ and $\rho^\star = -0.3$ (notice ρ is not used in the univariate case).

We first present results and discuss specific settings below. For the univariate g-and-k, Fig. 8 reports the marginal posterior distributions for each parameter at different number of observations for the considered methods. With increasing n, the BSL posterior does not concentrate (except for the parameter k); the energy score posterior concentrates close to the true value for all parameters (green vertical line), while the kernel score posterior performs slightly worse, not being able to concentrate for the parameter g (albeit this may happen with an even larger n, which we did not consider here). The poor performance of BSL is due to violation of the underlying normality assumption (which is to say, the scoring rule used by BSL is not strictly proper for this example), while the concentration of the energy and kernel score posteriors are in line with them being strictly proper SRs.

Similar results for the multivariate g-and-k are reported in Fig. 9. For this example, the PM-MCMCs targeting the semiBSL and BSL posteriors do not converge beyond respectively 1 and 10 observations; instead, with the Kernel and energy scores we do not experience such a problem. The energy score concentrates well on the exact parameter value in this case too, while the kernel score is able to concentrate well for some parameters (g and k) and some concentration can be observed for ρ; however, the kernel score posterior marginals

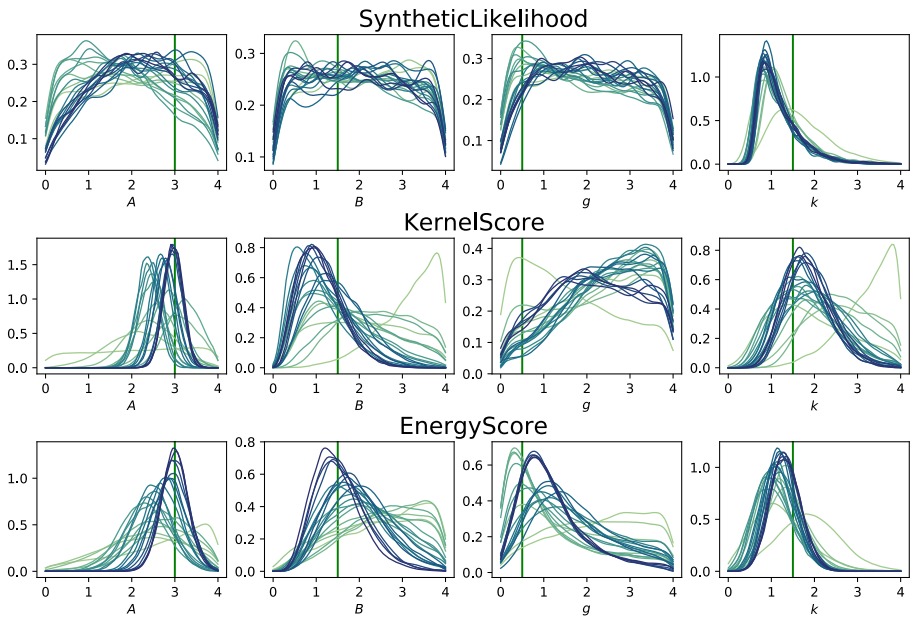


FIG 8. *Marginal posterior distributions for the different parameters for the well-specified univariate g-and-k model, with increasing number of observations ($n = 1, 5, 10, 15, \ldots, 100$), with PM-MCMC. Darker (respectively lighter) colors denote a larger (smaller) number of observations. The densities are obtained by KDE on the MCMC output thinned by a factor 10. The energy and kernel score posteriors concentrate around the true parameter value (green vertical line), while BSL does not.*

for A and B are flatter and noisier (it may be that larger n leads to more concentrate posterior for A and B as well, but we did not research this further).

We use the following settings for the SR posteriors:

- For the energy score posterior, our heuristic procedure (Sec. 4.2) for setting w using BSL as a reference resulted in $w \approx 0.35$ for the univariate model and $w \approx 0.16$ for the multivariate one.
- For the kernel score posterior, we first fit the value of the Gaussian kernel bandwidth parameter as described in Appendix E, which resulted in $\gamma \approx 5.50$ for the univariate case and $\gamma \approx 52.37$ for the multivariate one. Then, the heuristic procedure for w using BSL as a reference resulted in $w \approx 18.30$ for the univariate model and $w \approx 52.29$ for the multivariate one.

Next, we discuss the proposal sizes for PM-MCMC; recall that we use independent normal proposals on each component of θ, with standard deviation σ. We report here the values for σ used in the experiments; we stress that, as the PM-MCMC is run in the transformed unbounded parameter space (obtained applying a logit transformation), these proposal sizes refer to that space.

For the univariate g-and-k, the proposal sizes we use are the following:

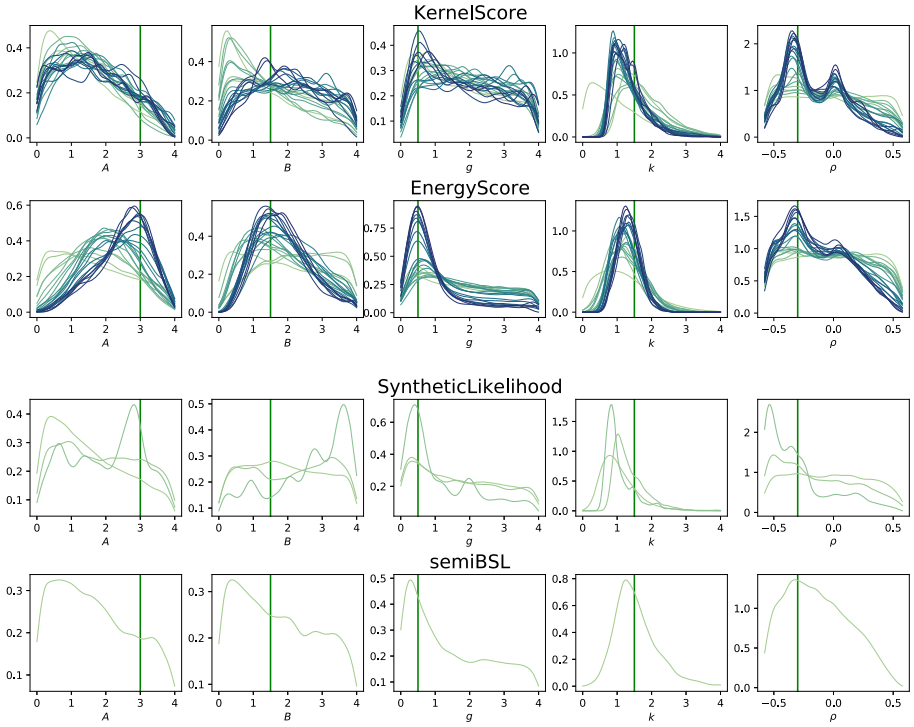


FIG 9. *Marginal posterior distributions for the different parameters for the well-specified multivariate g-and-k model, with increasing number of observations ($n = 1, 5, 10, 15, \ldots, 100$), with PM-MCMC. Darker (respectively lighter) colors denote a larger (smaller) number of observations. The densities are obtained by KDE on the MCMC output thinned by a factor 10. The energy score posterior concentrates well around the true parameter value (green vertical line), with the kernel score one performing slightly worse. For BSL, we were able to run the inference for $n = 1, 5, 10$, while we were only able to do so for $n = 1$ for semiBSL.*

- For BSL, we use $\sigma = 1$ for all values of n.
- For energy and kernel scores, we take $\sigma = 1$ for n from 1 up to 25 (included), $\sigma = 0.4$ for n from 30 to 50, and $\sigma = 0.2$ for n from 55 to 100.

For the multivariate g-and-k:

- For BSL and semiBSL, we use $\sigma = 1$ for all values of n for which the chain converges. We stress that we tried decreasing the proposal size, but that did not solve the non-convergence issue (discussed in the main text in Sec. 4.3.1).
- For energy and kernel scores, we take $\sigma = 1$ for n from 1 up to 15 (included), $\sigma = 0.4$ for n from 20 to 35, $\sigma = 0.2$ for n from 40 to 50 and $\sigma = 0.1$ for n from 55 to 100.

In Table 4, we report the acceptance rates the different methods achieve for all values of n, with the proposal sizes mentioned above. We denote by "/" the

TABLE 4
Acceptance rates for the univariate and multivariate g-and-k experiments with different values of n, with the PM-MCMC proposal sizes reported in Appendix G.1. "/" denotes experiments for which PM-MCMC did not run satisfactorily.

N. obs.	Univariate g-and-k			Multivariate g-and-k			
	BSL	Kernel score	Energy score	BSL	semiBSL	Kernel score	Energy score
1	0.362	0.507	0.420	0.216	0.190	0.468	0.445
5	0.221	0.329	0.375	0.069	/	0.136	0.224
10	0.133	0.252	0.272	0.036	/	0.127	0.216
15	0.109	0.253	0.217	/	/	0.077	0.154
20	0.100	0.154	0.207	/	/	0.151	0.278
25	0.092	0.149	0.208	/	/	0.126	0.233
30	0.085	0.218	0.343	/	/	0.124	0.222
35	0.080	0.172	0.315	/	/	0.076	0.166
40	0.076	0.152	0.293	/	/	0.119	0.246
45	0.070	0.130	0.256	/	/	0.103	0.223
50	0.062	0.121	0.220	/	/	0.103	0.219
55	0.060	0.189	0.317	/	/	0.139	0.297
60	0.059	0.185	0.324	/	/	0.129	0.286
65	0.057	0.173	0.314	/	/	0.133	0.273
70	0.052	0.172	0.289	/	/	0.119	0.256
75	0.048	0.161	0.273	/	/	0.123	0.247
80	0.048	0.159	0.267	/	/	0.117	0.233
85	0.045	0.150	0.252	/	/	0.098	0.213
90	0.044	0.143	0.247	/	/	0.087	0.198
95	0.044	0.136	0.244	/	/	0.089	0.198
100	0.042	0.129	0.236	/	/	0.076	0.190

experiments for which we did not manage to run PM-MCMC satisfactorily. We remark how the energy score achieves a larger acceptance rates in all experiments compared to the kernel score.

G.1.1. Investigating the poor PM-MCMC performance for BSL and semiBSL

The correlated pseudo-marginal MCMC for BSL and semiBSL performed poorly for the multivariate g-and-k example, not being able to converge when using more than respectively 1 and 10 observations We investigate now this poor performance, by fixing $n = 20$ and running PM-MCMC with 10 different initializations, for 10000 MCMC steps with no burn-in, for BSL and semiBSL, with $m = 500$. The chains look "sticky" and, after a short transient, get stuck in different regions of Θ (see Fig. 10).

In order to understand the reason for this result, we investigate whether the poor performance is due to large variance in the estimate of the target; as increasing the number of simulations reduces such variance, we study the effect of this on the PM-MCMC performance. Therefore, we report here the results of a study increasing the number of simulations for a fixed number of observations $n = 20$ for the g-and-k model. Specifically, we tested $m = 500, 1000, 1500, 2000, 2500, 3000, 30000$; as discussed in Appendix G.1, we used a proposal size $\sigma = 0.4$, with which the energy and kernel score posteriors performed well. We report traceplots in Fig. 11 and corresponding acceptance rates in Table 5; from this experiment, we note that BSL achieves acceptance rate

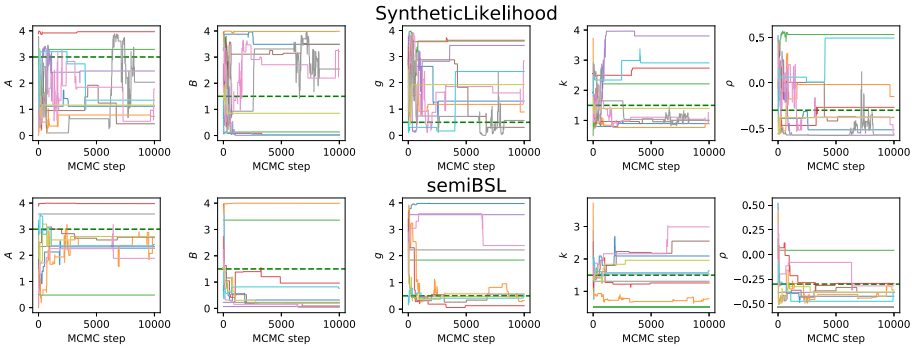


FIG 10. *Traceplots for semiBSL and BSL for $n = 20$ for 10 different initializations (different colors), with 10000 PM-MCMC steps (no burn-in); the green dashed line denotes the true parameter value. It can be seen that the chains are very sticky, and that they explore different parts of the parameter space.*

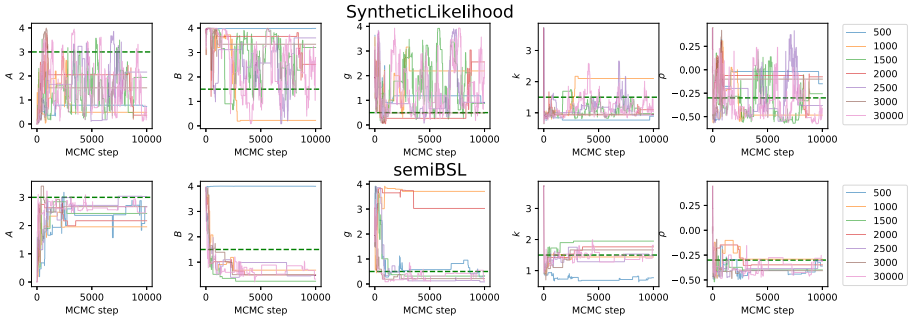


FIG 11. *Traceplots for BSL and semiBSL and BSL for $n = 20$ using different number of simulations m, reported in the legend for each row; green dashed line denotes the true parameter value. There is no improvement in the mixing of the chain for increasing the number of simulations.*

as large as few percentage points with larger m values, but there is no constant trend (for instance, acceptance rate with $m = 3000$ is smaller than with $m = 2000$), which means that the method is still prone to getting stuck. For semiBSL, the acceptance rate is abysmal even for very large m.

Additionally, while the BSL assumptions are unreasonable for this model, the multivariate g-and-k fulfills the assumptions underlying semiBSL: in fact, applying a one-to-one transformation to each component of a random vector does not change the copula structure, which is Gaussian in this case. It is therefore surprising that the performance of semiBSL degrades so rapidly when n increases.

TABLE 5
Acceptance rates for BSL and semiBSL and BSL for $n = 20$ using different number of simulations m; there is no improvement in the acceptance rate for increasing number of simulations. We recall that we were not able to run semiBSL for $m = 30000$ due to its high computational cost.

N. simulations m	500	1000	1500	2000	2500	3000	30000
Acc. rate BSL	$6.0 \cdot 10^{-3}$	$1.1 \cdot 10^{-2}$	$3.3 \cdot 10^{-2}$	$9.9 \cdot 10^{-3}$	$1.8 \cdot 10^{-2}$	$7.5 \cdot 10^{-3}$	$5.1 \cdot 10^{-2}$
Acc. rate semiBSL	$7.0 \cdot 10^{-3}$	$3.4 \cdot 10^{-3}$	$3.7 \cdot 10^{-3}$	$2.8 \cdot 10^{-3}$	$4.2 \cdot 10^{-3}$	$3.6 \cdot 10^{-3}$	$9.2 \cdot 10^{-3}$

G.2. Misspecified setup

The observations are here generated by a Cauchy distribution. For the univariate case, the univariate Cauchy is used; for the multivariate case, the observations are generated as in Sec. 4.3.1 (i.e., no correlation between components).

In order to have coherent results with respect to the well-specified case, we use here the values of w and γ determined in the well-specified case (reported in Appendix G.1)

For the univariate g-and-k, we report the marginal posteriors in Fig. 12. The energy and kernel score posteriors concentrate on a similar parameter value; the BSL posterior concentrates as well (differently from the well-specified case), albeit on a slightly different parameter value (especially for B and k). Therefore, with this kind of misspecification, $\theta^\star$ is unique both when using the strictly proper Kernel and energy scores, as well as the non-strictly proper Dawid–Sebastiani Score (corresponding to BSL).

For the multivariate g-and-k, we experienced the same issue with PM-MCMC as in the well-specified case for BSL and semiBSL; therefore, we do not report those results. Marginals for the energy and kernel score posteriors can be seen in Fig. 13; both posteriors concentrate for all parameters except for ρ (which describes correlation among different components in the observations, here absent). For the other parameters, the two methods concentrate on very similar parameter values, with slightly larger difference for k, for which the kernel score posterior does not concentrate very well.

The above resuts are obtained with the following proposal sizes for PM-MCMC (which is run with independent normal proposals on each component of θ with standard deviation σ, in the same way as in the well-specified case, after applying a logit transformation to the parameter space).

- For the univariate g-and-k, for all methods (BSL, energy and kernel scores), we take $\sigma = 1$ for n from 1 up to 25 (included), $\sigma = 0.4$ for n from 30 to 50, and $\sigma = 0.2$ for n from 55 to 100.
- For the multivariate g-and-k, recall that we did not report results for BSL and semiBSL here as we were not able to sample the posteriors with PM-MCMC for large n, as already experienced in the well-specified case. For the remaining techniques, we used the same values of σ as in the well-specified experiments (Appendix F.1.2).

In Table 6, we report the acceptance rates the different methods achieve for

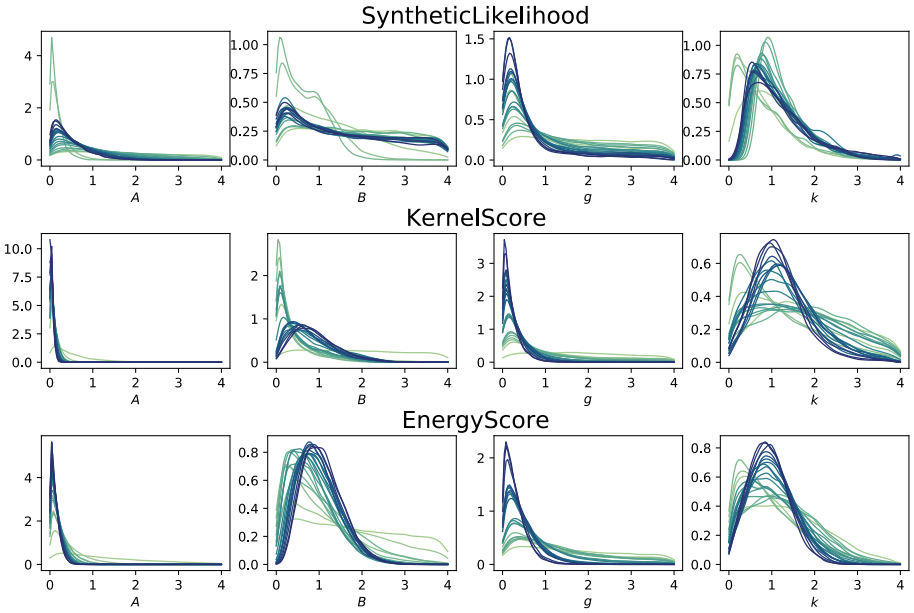


FIG 12. *Marginal posterior distributions for the different parameters for the univariate g-and-k model, with increasing number of observations ($n = 1, 5, 10, 15, \ldots, 100$) generated from the Cauchy distribution, with PM-MCMC. Darker (respectively lighter) colors denote a larger (smaller) number of observations. The densities are obtained by KDE on the MCMC output thinned by a factor 10. The energy and kernel score posteriors concentrate around the same parameter value, while BSL concentrates on slightly different one (specially for B and k).*

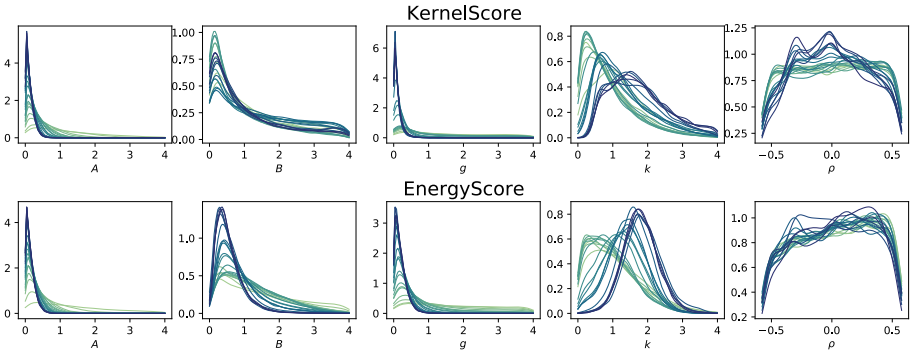


FIG 13. *Marginal posterior distributions for the different parameters for the multivariate g-and-k model, with increasing number of observations ($n = 1, 5, 10, 15, \ldots, 100$) generated from the Cauchy distribution, with PM-MCMC. Darker (respectively lighter) colors denote a larger (smaller) number of observations The densities are obtained by KDE on the MCMC output thinned by a factor 10. Both energy and kernel score posteriors concentrate on a very similar parameter value, with slightly larger difference for k.*

all values of n, with the proposal sizes discussed above. We remark how the energy score achieves a larger acceptance rates in all experiments compared to the kernel score.

TABLE 6
Acceptance rates for the misspecified univariate and multivariate g-and-k experiments with different values of n, with the PM-MCMC proposal sizes reported in Appendix G.2.

N. obs.	Misspecified univariate g-and-k			Misspecified multivariate g-and-k	
	BSL	Kernel score	Energy score	Kernel score	Energy score
1	0.457	0.482	0.521	0.472	0.470
5	0.302	0.436	0.454	0.324	0.373
10	0.193	0.450	0.425	0.362	0.330
15	0.146	0.441	0.390	0.361	0.276
20	0.102	0.264	0.311	0.544	0.410
25	0.093	0.288	0.314	0.530	0.377
30	0.153	0.426	0.471	0.536	0.359
35	0.144	0.349	0.448	0.537	0.336
40	0.134	0.340	0.440	0.631	0.432
45	0.130	0.344	0.429	0.523	0.373
50	0.125	0.255	0.393	0.383	0.343
55	0.167	0.318	0.501	0.471	0.436
60	0.176	0.303	0.490	0.412	0.407
65	0.164	0.293	0.481	0.389	0.391
70	0.164	0.276	0.455	0.372	0.374
75	0.156	0.272	0.445	0.278	0.329
80	0.157	0.262	0.436	0.232	0.306
85	0.153	0.254	0.430	0.247	0.300
90	0.147	0.231	0.415	0.239	0.299
95	0.152	0.226	0.410	0.235	0.291
100	0.141	0.223	0.407	0.232	0.277

Appendix H: Effect of m on pseudo-marginal MCMC

Here, we consider the univariate and multivariate g-and-k, both well-specified and misspecified, and study the impact of varying m in the resulting PM-MCMC target. As we span from very small to large values of m, we use here the vanilla pseudo-marginal MCMC of [4] instead of the correlated pseudo-marginal MCMC which was used for all other simulations. In this Appendix, we only consider the summary-free version of BSL (see Section 4.3.1).

The choice of m has two different impacts on the PM-MCMC:

1. first, it changes the pseudo-marginal MCMC target, as discussed in Section 3 in the main text; recall how, there, we proved that, for $m \to \infty$, the pseudo-marginal MCMC target converges to the original SR posterior defined in Eq. (2) in the main text. Therefore, we expect, for large enough m, the pseudo-marginal MCMC target to be roughly constant.
2. Additionally, smaller values of m imply that the target estimate has a larger variance. Therefore, we expect sampling to be harder for small m,

in terms of acceptance rate of the MCMC, and easier for large m (albeit that is more computationally intensive).

In our simulation study below, we consider m values from 10 to 1000. Our results empirically verify our expectations above. In particular, we find that, for m larger than a threshold which is typically few hundreds, the pseudo-marginal MCMC target is roughly constant. Additionally, very small values of m (few tens) make sampling impractical.

Moreover, our empirical results suggest that larger values of m are required for the PM-MCMC for semiBSL to be stable. For the other methods, the required m seem to be fairly similar, with slightly larger values for BSL for some models.

Typically, we found m values in the few hundreds to strike a good balance between larger computational cost and improved acceptance rate with larger m. Additionally, this consideration depends also on how quickly the simulation cost scales with m: even when not parallelizing model simulations across different processors, if the implementation is vectorized, the computational cost can scale sub-linearly in m, which means a better PM-MCMC efficiency is reached for a larger m. A more extensive study considering for instance the effective sample size per CPU time could be carried out.

In all experiments, except where said otherwise, we use the value of w found via our heuristics strategy (Section 4.2 in the main text) and reported above.

H.1. Univariate g-and-k

Here, we report results considering $n = 10$ observations.

TABLE 7

Acceptance rate and trace of the posterior covariance matrix for different values of m for the well-specified univariate g-and-k, for the BSL, Kernel and energy score posteriors.

m	**BSL**		**Kernel score**		**Energy score**	
	Acc. rate	$\mathrm{Tr}\,[\Sigma_{\mathrm{post}}]$	Acc. rate	$\mathrm{Tr}\,[\Sigma_{\mathrm{post}}]$	Acc. rate	$\mathrm{Tr}\,[\Sigma_{\mathrm{post}}]$
10	0.104	4.5245	0.011	3.6030	0.063	3.9822
20	0.122	4.4439	0.035	3.6679	0.115	3.9642
50	0.129	4.3778	0.098	3.3803	0.179	3.6105
100	0.134	4.4095	0.157	3.2220	0.219	3.5335
200	0.136	4.1753	0.204	3.1628	0.243	3.4730
300	0.135	4.2261	0.220	3.1181	0.252	3.3537
400	0.135	4.1769	0.229	3.0716	0.257	3.3553
500	0.132	4.1702	0.234	3.1079	0.262	3.4362
600	0.130	4.2095	0.239	3.0295	0.259	3.2612
700	0.133	4.2417	0.243	3.0536	0.265	3.3629
800	0.132	4.2421	0.247	3.0216	0.265	3.3077
900	0.132	4.1084	0.248	3.0477	0.267	3.3815
1000	0.137	4.2930	0.253	3.1181	0.269	3.3570

H.2. Misspecified univariate g-and-k

Here, we report results considering $n = 10$ observations.

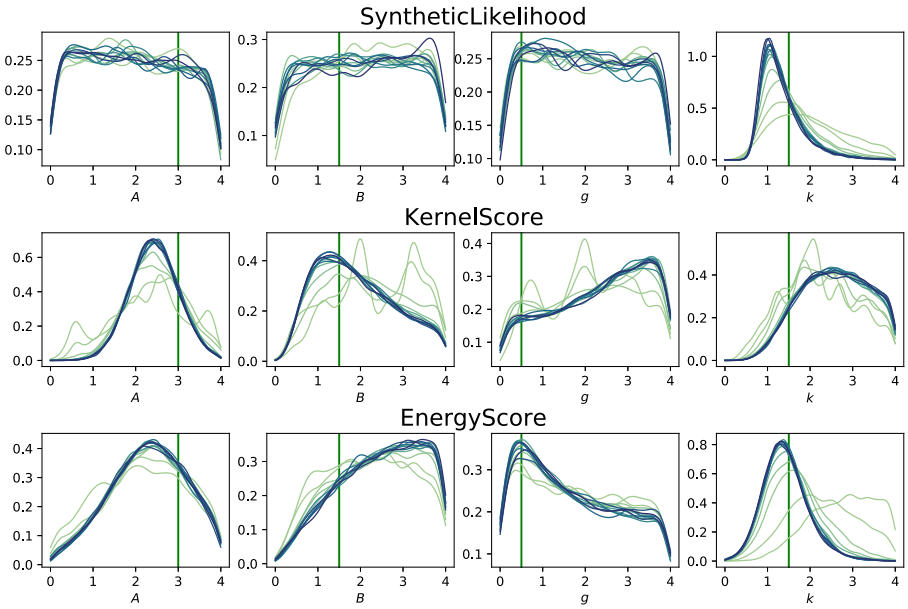


FIG 14. *Univariate posterior marginals for different m values for the well-specified univariate g-and-k distribution, for the BSL, Kernel and energy score posteriors, with PM-MCMC. Lighter (respectively darker) colors denote smaller (resp. larger) values of m. For small values of m, the marginals are spiky, which is due to unstable PM-MCMC. The densities are obtained by KDE on the MCMC output thinned by a factor 10.*

TABLE 8

Acceptance rate and trace of the posterior covariance matrix for different values of m for the misspecified univariate g-and-k, for the BSL, Kernel and energy score posteriors.

m	**BSL**		**Kernel score**		**Energy score**	
	Acc. rate	$\mathrm{Tr}\left[\Sigma_{\mathrm{post}}\right]$	Acc. rate	$\mathrm{Tr}\left[\Sigma_{\mathrm{post}}\right]$	Acc. rate	$\mathrm{Tr}\left[\Sigma_{\mathrm{post}}\right]$
10	0.038	3.3664	0.047	3.4141	0.164	3.8095
20	0.072	2.3207	0.069	3.2060	0.216	3.4900
50	0.130	1.9729	0.184	2.6690	0.306	2.9483
100	0.159	2.0145	0.298	2.4529	0.364	2.7232
200	0.179	1.8829	0.359	2.4037	0.391	2.7153
300	0.187	2.0198	0.389	2.3623	0.402	2.6055
400	0.188	1.9498	0.405	2.3403	0.410	2.6164
500	0.189	1.9092	0.412	2.3756	0.413	2.5579
600	0.191	1.8259	0.422	2.3461	0.414	2.5704
700	0.186	1.9207	0.430	2.3452	0.417	2.5484
800	0.184	1.9509	0.432	2.3810	0.419	2.6276
900	0.190	1.9475	0.434	2.4472	0.423	2.6468
1000	0.194	1.9763	0.436	2.3434	0.425	2.6386

H.3. Multivariate g-and-k

Here, we report results considering $n = 10$ observations.

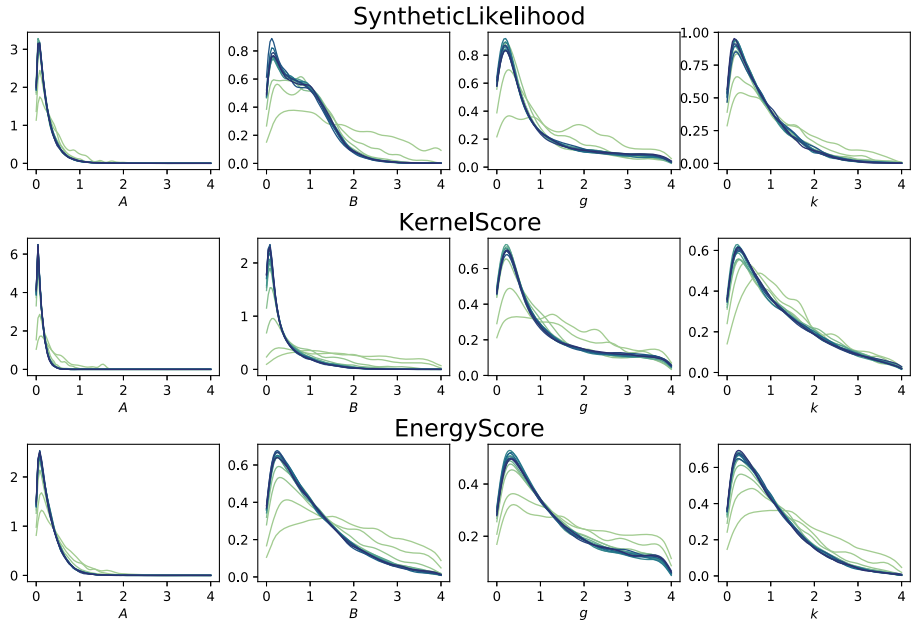


FIG 15. *Univariate posterior marginals for different m values for the misspecified univariate g-and-k distribution, for the BSL, Kernel and energy score posteriors, with PM-MCMC. Lighter (respectively darker) colors denote smaller (resp. larger) values of m. The densities are obtained by KDE on the MCMC output thinned by a factor 10.*

TABLE 9

Acceptance rate and trace of the posterior covariance matrix for different values of m for the well-specified multivariate g-and-k, for the BSL, semiBSL, Kernel and energy score posteriors.

m	**BSL**		**semiBSL**		**Kernel score**		**Energy score**	
	Acc. rate	Tr $[\Sigma_{\text{post}}]$	Acc. rate	Tr $[\Sigma_{\text{post}}]$	Acc. rate	Tr $[\Sigma_{\text{post}}]$	Acc. rate	Tr $[\Sigma_{\text{post}}]$
10	<0.001	1.0566	<0.001	0.4227	0.006	3.6061	0.070	4.5255
20	<0.001	0.3674	<0.001	0.6383	0.023	4.0455	0.123	3.9212
50	0.003	2.8320	<0.001	0.6331	0.055	3.8924	0.170	3.8571
100	0.002	2.3666	<0.001	0.6131	0.078	4.1250	0.194	3.8126
200	0.001	0.7140	0.001	0.8603	0.099	3.9624	0.206	3.7142
300	0.008	2.8229	0.002	2.2184	0.108	4.2766	0.208	3.9078
400	0.009	2.5694	0.001	0.6885	0.113	3.9710	0.212	3.8284
500	0.009	3.3583	0.002	1.2885	0.116	4.0250	0.217	3.8383
600	0.013	2.9646	0.005	1.3359	0.120	3.9632	0.216	3.7698
700	0.010	3.7043	0.005	0.6511	0.119	4.0173	0.214	3.7437
800	0.016	3.3017	0.006	0.6679	0.122	3.9607	0.214	3.7512
900	0.022	2.9915	0.005	0.6411	0.126	4.1293	0.216	3.9202
1000	0.017	3.1304	0.006	0.5892	0.122	3.9757	0.216	3.7959

For this model, small m lead to extremely small acceptance rates for BSL and semiBSL (Table 9); in those cases, the trace of the posterior covariance matrix is also very small due to the chain being almost still. Additionally, even large m values lead to small acceptance rate for semiBSL; that is consequence of the

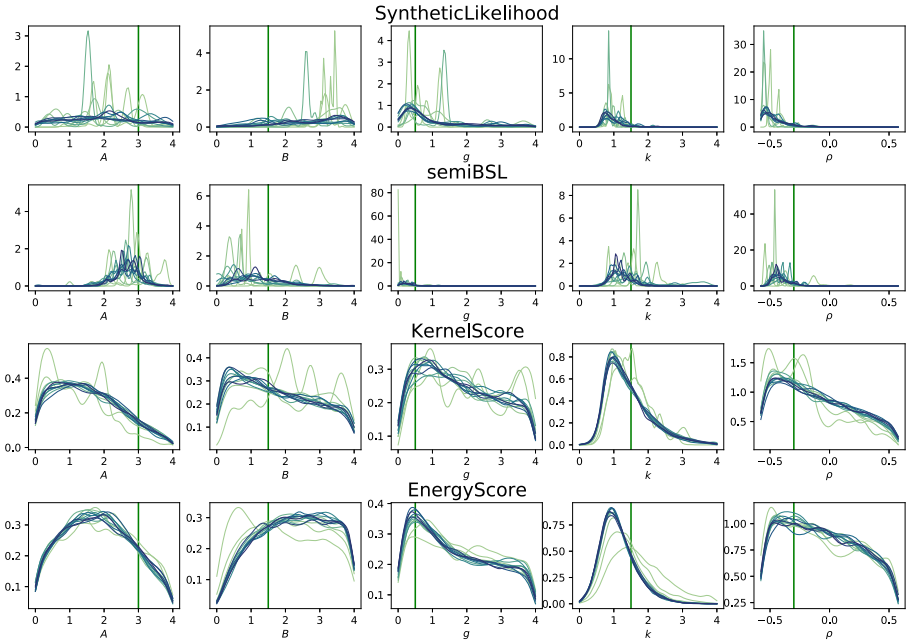


FIG 16. *Univariate posterior marginals for different m values for the well-specified multivariate g-and-k distribution, for the BSL, semiBSL, Kernel and energy score posteriors, with PM-MCMC. Lighter (respectively darker) colors denote smaller (resp. larger) values of m. For small values of m, the marginals are spiky, which is due to unstable MCMC. The densities are obtained by KDE on the MCMC output thinned by a factor 10.*

TABLE 10

Acceptance rate and trace of the posterior covariance matrix for different values of m for the misspecified multivariate g-and-k, for the Kernel and energy score posteriors.

m	**Kernel score**		**Energy score**	
	Acc. rate	Tr $[\Sigma_{\text{post}}]$	Acc. rate	Tr $[\Sigma_{\text{post}}]$
10	0.017	4.5045	0.174	3.4306
20	0.108	3.6950	0.252	3.2373
50	0.243	3.4612	0.300	3.0291
100	0.308	3.4759	0.316	3.0081
200	0.344	3.4666	0.323	2.9303
300	0.348	3.4583	0.321	2.9160
400	0.355	3.4158	0.331	3.0031
500	0.359	3.4047	0.332	2.9743
600	0.363	3.3847	0.330	2.9321
700	0.360	3.3485	0.329	2.9249
800	0.361	3.3505	0.332	2.9854
900	0.363	3.3627	0.331	3.0155
1000	0.363	3.3307	0.330	2.9277

issues discussed in Appendix G.1.1. We report nevertheless the results here.

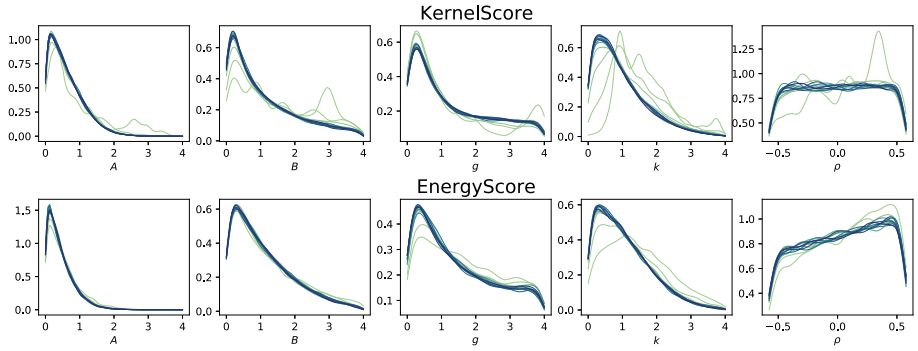


FIG 17. *Univariate posterior marginals for different m values for the misspecified multivariate g-and-k distribution, for the Kernel and energy score posteriors, with PM-MCMC. Lighter (respectively darker) colors denote smaller (resp. larger) values of m. For small values of m, the marginals are spiky, which is due to unstable MCMC. The densities are obtained by KDE on the MCMC output thinned by a factor 10.*

H.4. Misspecified multivariate g-and-k

Here, we report results considering $n = 10$ observations. We do not report results for BSL and semiBSL as those were unable to run satisfactorily for that number of observations, for all considered values of m.

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments that improved the quality of this paper. We thank Jeremias Knoblauch, François-Xavier Briol, Takuo Matsubara, Geoff Nicholls, Benedict Leimkuhler and Sebastian Schmon for valuable feedback and suggestions on earlier versions of this work. We also thank Alex Shestopaloff for providing code for exact MCMC for the M/G/1 model.
Lorenzo Pacchiardi conducted part of this work during his PhD studies at the Department of Statistics, University of Oxford. Sherman Khoo conducted part of this work during his MSc studies at the Department of Statistics, University of Warwick.

Funding

LP was supported by the EPSRC and MRC through the OxWaSP CDT programme (EP/L016710/1) during his PhD, which also funded part of the computational resources used to perform this work. LP is currently funded by US DARPA grant HR00112120007 (RECoG-AI). SK is supported by the EPSRC (grant number EP/S023569/1).
RD is funded by EPSRC (grant nos. EP/V025899/1, EP/T017112/1) and NERC (grant no. NE/T00973X/1).

Supplementary Material

Code

(doi: 10.1214/24-EJS2283SUPP; .zip). Code necessary to reproduce all experiments.

References

[1] An, Z., Nott, D. J. and Drovandi, C. (2020). Robust Bayesian synthetic likelihood via a semi-parametric approach. *Statistics and Computing* **30** 543–557. MR4065218

[2] An, Z., South, L. F. and Drovandi, C. (2019). BSL: An R package for efficient parameter estimation for simulation-based models via Bayesian synthetic likelihood. *arXiv preprint arXiv:1907.10940*.

[3] An, Z., South, L. F., Nott, D. J. and Drovandi, C. C. (2019). Accelerating Bayesian synthetic likelihood with the graphical lasso. *Journal of Computational and Graphical Statistics* **28** 471–475. MR3974895

[4] Andrieu, C., Roberts, G. O. et al. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics* **37** 697–725. MR2502648

[5] Arnold, H., Moroz, I. and Palmer, T. (2013). Stochastic parametrizations and model uncertainty in the Lorenz'96 system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **371** 20110479.

[6] 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. https://doi.org/10.1111/rssb.12312 MR3928142

[7] Bhattacharya, A., Pati, D. and Yang, Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics* **47** 39–66. MR3909926

[8] Billingsley, P. (1999). *Convergence of probability measures*, 2nd ed. John Wiley & Sons. MR1700749

[9] Bińkowski, M., Sutherland, D. J., Arbel, M. and Gretton, A. (2018). Demystifying MMD GANs. In *International Conference on Learning Representations*.

[10] Bissiri, P. G., Holmes, C. C. and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B, Statistical methodology* **78** 1103. MR3557191

[11] Boudt, K., Cornelissen, J. and Croux, C. (2012). The Gaussian rank correlation estimator: robustness properties. *Statistics and Computing* **22** 471–483. MR2865030

[12] Briol, F.-X., Barp, A., Duncan, A. B. and Girolami, M. (2019). Statistical inference for generative models with Maximum Mean Discrepancy. *arXiv preprint arXiv:1906.05944*. MR3577382

[13] Chérief-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes: Robust Bayesian Estimation via Maximum Mean Discrepancy. In *Symposium on Advances in Approximate Bayesian Inference* 1–21. PMLR. MR4147569

[14] Chwialkowski, K., Strathmann, H. and Gretton, A. (2016). A kernel test of goodness of fit. In *International conference on machine learning* 2606–2615. PMLR.

[15] Corbella, A., Spencer, S. E. and Roberts, G. O. (2022). Automatic Zig-Zag sampling in practice. *Statistics and Computing* **32** 107. MR4507158

[16] Coullon, J., South, L. and Nemeth, C. (2023). Efficient and generalizable tuning strategies for stochastic gradient MCMC. *Statistics and Computing* **33** 66. MR4572153

[17] Dahlin, J., Lindsten, F., Kronander, J. and Schön, T. B. (2015). Accelerating pseudo-marginal Metropolis-Hastings by correlating auxiliary variables. *arXiv preprint arXiv:1511.05483*.

[18] Dawid, A. P. and Musio, M. (2014). Theory and applications of proper scoring rules. *Metron* **72** 169–183. MR3233147

[19] Del Moral, P., Doucet, A. and Jasra, A. (2012). An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing* **22** 1009–1020. MR2950081

[20] Deligiannidis, G., Doucet, A. and Pitt, M. K. (2018). The correlated pseudomarginal method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 839–870. MR3874301

[21] Dellaporta, C., Knoblauch, J., Damoulas, T. and Briol, F.-X. (2022). Robust Bayesian Inference for Simulator-based Models via the MMD Posterior Bootstrap. In *International Conference on Artificial Intelligence and Statistics* 943–970. PMLR.

[22] Ding, N., Fang, Y., Babbush, R., Chen, C., Skeel, R. D. and Neven, H. (2014). Bayesian sampling using stochastic gradient thermostats. *Advances in neural information processing systems* **27**.

[23] Drovandi, C. C. and Pettitt, A. N. (2011). Likelihood-free Bayesian estimation of multivariate quantile distributions. *Computational Statistics & Data Analysis* **55** 2541–2556. MR2802334

[24] Drovandi, C. C., Pettitt, A. N. and Lee, A. (2015). Bayesian Indirect Inference Using a Parametric Auxiliary Model. *Statistical Science* **30** 72–95. MR3317755

[25] Duffield, S., Benedetti, M. and Rosenkranz, M. (2023). Bayesian learning of parameterised quantum circuits. *Machine Learning: Science and Technology* **4** 025007.

[26] Dutta, R., Schoengens, M., Pacchiardi, L., Ummadisingu, A., Widmer, N., Künzli, P., Onnela, J.-P. and Mira, A. (2021). ABCpy: A High-Performance Computing Perspective to Approximate Bayesian Computation. *Journal of Statistical Software* **100** 1–38. https://doi.org/10.18637/jss.v100.i07

[27] Fearnhead, P., Bierkens, J., Pollock, M. and Roberts, G. O. (2018). Piecewise deterministic Markov processes for continuous-time

Monte Carlo. *Statistical Science* **33** 386–412. MR3843382

[28] Fong, E., Lyddon, S. and Holmes, C. (2019). Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *International Conference on Machine Learning* 1952–1962. PMLR.

[29] Frazier, D. T., Kohn, R., Drovandi, C. and Gunawan, D. (2023). Reliable Bayesian Inference in Misspecified Models. *arXiv preprint arXiv: 2302.06031*.

[30] Frazier, D. T., Martin, G. M., Robert, C. P. and Rousseau, J. (2018). Asymptotic properties of approximate Bayesian computation. *Biometrika* **105** 593–607. MR3842887

[31] Frazier, D. T., Nott, D. J., Drovandi, C. and Kohn, R. (2022). Bayesian Inference Using Synthetic Likelihood: Asymptotics and Adjustments. *Journal of the American Statistical Association* **0** 1-12. https://doi.org/10.1080/01621459.2022.2086132 MR4681623

[32] Frazier, D. T., Robert, C. P. and Rousseau, J. (2020). Model misspecification in approximate Bayesian computation: consequences and diagnostics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82** 421-444. https://doi.org/10.1111/rssb.12356 MR4084170

[33] Ghosh, J. K., Delampady, M. and Samanta, T. (2006). *An introduction to Bayesian analysis: theory and methods* **725**. Springer. MR2247439

[34] Giummolè, F., Mameli, V., Ruli, E. and Ventura, L. (2019). Objective Bayesian inference with proper scoring rules. *Test* **28** 728–755. MR3992136

[35] Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association* **102** 359–378. MR2345548

[36] Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *International Conference on Machine Learning* 1292–1301. PMLR.

[37] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research* **13** 723–773. MR2913716

[38] Hakkarainen, J., Ilin, A., Solonen, A., Laine, M., Haario, H., Tamminen, J., Oja, E. and Järvinen, H. (2012). On closure parameter estimation in chaotic systems. *Nonlinear processes in Geophysics* **19** 127–143.

[39] Holmes, C. and Walker, S. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika* **104** 497–503. MR3698270

[40] Jarvenpaa, M., Vehtari, A. and Marttinen, P. (2020). Batch simulations and uncertainty quantification in Gaussian process surrogate approximate Bayesian computation. In *Conference on Uncertainty in Artificial Intelligence* 779–788. PMLR.

[41] Jewson, J., Smith, J. Q. and Holmes, C. (2018). Principles of Bayesian inference using general divergence criteria. *Entropy* **20** 442. MR3879894

[42] Jiang, B. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *International Conference on Artificial Intelligence and Statistics* 1711–1721.

[43] Jones, A. and Leimkuhler, B. (2011). Adaptive stochastic methods for sampling driven molecular systems. *The Journal of chemical physics* **135** 084125.

[44] Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.).

[45] Knoblauch, J., Jewson, J. and Damoulas, T. (2022). An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference. *Journal of Machine Learning Research* **23** 1–109. MR4577084

[46] Leimkuhler, B., Sachs, M. and Stoltz, G. (2020). Hypocoercivity properties of adaptive Langevin dynamics. *SIAM Journal on Applied Mathematics* **80** 1197–1222. MR4099815

[47] Leimkuhler, B. and Shang, X. (2016). Adaptive thermostats for noisy gradient systems. *SIAM Journal on Scientific Computing* **38** A712–A736. MR3465428

[48] Li, C., Chen, C., Carlson, D. and Carin, L. (2016). Preconditioned stochastic gradient Langevin dynamics for deep neural networks. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[49] Li, W. and Fearnhead, P. (2018). Convergence of regression-adjusted approximate Bayesian computation. *Biometrika* **105** 301–318. MR3804404

[50] Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S. and Corander, J. (2017). Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology* **66** e66–e82. https://doi.org/10.1093/sysbio/syw077

[51] Liu, Q., Lee, J. and Jordan, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *International conference on machine learning* 276–284. PMLR.

[52] Loaiza-Maya, R., Martin, G. M. and Frazier, D. T. (2021). Focused Bayesian prediction. *Journal of Applied Econometrics* **36** 517–543. MR4309597

[53] Lorenz, E. N. (1996). Predictability: A problem partly solved. In *Proc. Seminar on predictability* **1**. MR1210968

[54] Lyddon, S., Holmes, C. and Walker, S. (2019). General Bayesian updating and the loss-likelihood bootstrap. *Biometrika* **106** 465–478. MR3949315

[55] Lyddon, S., Walker, S. and Holmes, C. C. (2018). Nonparametric learning from Bayesian models with randomized objective functions. *Advances in Neural Information Processing Systems* **31**.

[56] Matsubara, T., Knoblauch, J., Briol, F.-X. and Oates, C. J. (2022). Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **84** 997-1022. https://doi.org/10.1111/rssb.12500 MR4460583

[57] Matsubara, T., Knoblauch, J., Briol, F.-X. and Oates, C. J. (2023). Generalized Bayesian Inference for Discrete Intractable Likelihood. *Journal*

of the American Statistical Association 1–11. MR4460583

[58] Miller, J. W. (2021). Asymptotic Normality, Concentration, and Coverage of Generalized Posteriors. *Journal of Machine Learning Research* **22** 1–53. MR4318524

[59] Nemeth, C. and Fearnhead, P. (2021). Stochastic gradient markov chain monte carlo. *Journal of the American Statistical Association* **116** 433–450. MR4227705

[60] Nguyen, H. D., Arbel, J., Lü, H. and Forbes, F. (2020). Approximate Bayesian computation via the energy statistic. *IEEE Access* **8** 131683–131698.

[61] Pacchiardi, L. (2022). Statistical inference in generative models using scoring rules, PhD thesis, University of Oxford. MR4716183

[62] Pacchiardi, L. and Dutta, R. (2022). Score Matched Neural Exponential Families for Likelihood-Free Inference. *Journal of Machine Learning Research* **23** 1-71. MR4420763

[63] Pacchiardi, L. and Khoo, S. and Dutta, R. (2024). Supplement to "Generalized Bayesian likelihood-free inference".

[64] Pagani, F., Chevallier, A., Power, S., House, T. and Cotter, S. (2024). NuZZ: Numerical Zig-Zag for general models. *Statistics and Computing* **34** 61. MR4686092

[65] Park, M., Jitkrittum, W. and Sejdinovic, D. (2016). K2-ABC: Approximate Bayesian computation with kernel embeddings. In *Artificial Intelligence and Statistics*.

[66] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox and R. Garnett, eds.) 8024–8035. Curran Associates, Inc.

[67] Pauli, F., Racugno, W. and Ventura, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica* 149–164. MR2796857

[68] Picchini, U., Simola, U. and Corander, J. (2022). Sequentially Guided MCMC Proposals for Synthetic Likelihoods and Correlated Synthetic Likelihoods. *Bayesian Analysis* 1 – 31. https://doi.org/10.1214/22-BA1305 MR4674633

[69] Pillai, N. S., Stuart, A. M. and Thiéry, A. H. (2012). Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. MR3024970

[70] Prangle, D. (2017). gk: An R Package for the g-and-k and Generalised g-and-h Distributions. *arXiv preprint arXiv:1706.06889*.

[71] Price, L. F., Drovandi, C. C., Lee, A. and Nott, D. J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics* **27** 1–11. MR3788296

[72] Rizzo, M. L. and Székely, G. J. (2016). Energy distance. *Wiley inter-*

disciplinary reviews: Computational statistics **8** 27–38. MR3457239

[73] Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* 341–363. MR1440273

[74] Ruli, E., Sartori, N. and Ventura, L. (2016). Approximate Bayesian computation with composite score functions. *Statistics and Computing* **26** 679–692. MR3489864

[75] Scheffé, H. (1947). A useful convergence theorem for probability distributions. *The Annals of Mathematical Statistics* **18** 434–438. MR0021585

[76] Syring, N. and Martin, R. (2019). Calibrating general posterior credible regions. *Biometrika* **106** 479–486. MR3949316

[77] Thomas, O., Dutta, R., Corander, J., Kaski, S., Gutmann, M. U. et al. (2020). Likelihood-free inference by ratio estimation. *Bayesian Analysis*. MR4377135

[78] Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)* 681–688.

[79] Wilks, D. S. (2005). Effects of stochastic parametrizations in the Lorenz'96 system. *Quarterly Journal of the Royal Meteorological Society* **131** 389–407.

[80] Wilks, D. S. (2019). Chapter 9 – Forecast Verification. In *Statistical Methods in the Atmospheric Sciences* Fourth ed. (D. S. Wilks, ed.) 369-483. Elsevier. https://doi.org/10.1016/B978-0-12-815823-4.00009-2