Electronic Journal of Statistics
Vol. 18 (2024) 5156–5197
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2324

Approximate Bayesian computation using the Fourier integral theorem

Frank Rotiroti[1] **and Stephen G. Walker**[2]

[1] *Department of Statistics and Data Sciences, UT Austin*
e-mail: rotiroti@utexas.edu

[2] *Department of Mathematics, UT Austin*
e-mail: s.g.walker@math.utexas.edu

Abstract: We present a general method for obtaining posterior samples from a model for which the likelihood function is intractable or otherwise unavailable, but from which simulation is possible. While the inability to evaluate the likelihood impedes the implementation of traditional sampling algorithms, like the Metropolis–Hastings algorithm, our approach, based on the Fourier integral theorem, allows for an exact simulation-based estimate of such likelihoods. Moreover, given its foundations in the Fourier integral theorem, our approach comes with clear guidelines for setting the two tuning parameters on which it relies. Through comparison with alternative methods, specifically synthetic likelihoods and ABC, including both dimension-reduction and full-data approaches, across a variety of examples, culminating in a highly nonlinear state space model based on the Ricker map, we will demonstrate how the Fourier approach is able to provide statistically sensible, full-data likelihood estimation, requiring only that data can be simulated from the model.

MSC2020 subject classifications: Primary 62G07, 62F15.
Keywords and phrases: Fourier integral theorem, intractable likelihood, approximate Bayesian computation, simulation-based modeling.

Received November 2023.

1. Introduction

Arising in a variety of disciplines, including ecology, genetics, and finance, statistical models with intractable likelihood functions present a challenge for estimation. The approximate Bayesian computation (ABC) assumption is that while the likelihood is unavailable to be enumerated, it is possible to sample outcomes from the corresponding probability model for all sets of parameter values. The original ABC approach was not actually approximate at all; if the data are discrete and a parameter is chosen from the prior and then a new data set of equivalent size to the observed data is sampled using this parameter, the parameter can be taken as an output from the posterior if the sampled data coincide with the observed data; see [24].

Current approaches to ABC are aimed at implementing a similar strategy when observations are continuous. There are by now a wide variety of ideas, also known as likelihood-free techniques, such as those surveyed by [17], as

well as synthetic likelihood (SL) [30] and full-data distance-based approaches [10].

ABC and SL, in practice, rely on information reduction. For ABC, the data are typically reduced to a set of summary statistics, for which an associated tolerance and distance must also be specified. Calibrating these tuning parameters can be difficult, with effects that can prove unpredictable. For SL, the approximation of the likelihood is underpinned by a distributional assumption of approximate normality on the vector of summary statistics. In addition to this assumption of normality, as well as the need to determine a suitable set of summary statistics, SL must also estimate the covariance matrix for a normal distribution. As models become more complicated, because of hierarchical structure or time dependence, these challenges become more acute.

Full-data distance-based approaches seek to avoid the reduction of the data to a set of summary statistics by comparing the (marginal) distributions of the observed and simulated data according to some measure of distance, such as the Wasserstein distance or Kullback-Leibler divergence. Although these approaches use the full data, the application of a single distance function may nonetheless result in information loss, as discussed in [10]. Accordingly, in the implementation of such approaches, it might be necessary to introduce summary statistics or to combine multiple distance functions in a weighted average. Large or dependent data sets tend to exacerbate these issues, and the procedures to address them can be highly problem dependent, thus hampering the general feasibility of these approaches.

When faced with an intractable likelihood, one might still attempt to estimate the likelihood function. The concern then shifts to the level of approximation. We treat the problem of the estimation of the likelihood as one of estimating a density at a point or a series of multiple points. Such an approach has recently been used for estimating a posterior density at a point; see [23].

The present paper extends [23], who applied a Fourier approach to estimating marginal likelihoods, in a number of significant ways. In the marginal likelihood setting, the point at which the density is estimated can be chosen to be any high-density point in the support of the posterior. In the ABC setting, the points are given, since they are the data, with the potential for low-density points, thus requiring additional transformations and more specific theory as to the behavior of the estimator for points in the tails of a distribution.

Also, in the marginal likelihood setting, the number of samples on which the density estimation is to be based can typically be taken to be arbitrarily large. In the ABC setting, the likelihood estimation is used repeatedly at each iteration of an MCMC sampler, thus requiring a more thorough examination of the behavior of the estimator for smaller sample sizes.

Thus, we have generally worked with smaller values of a Monte Carlo sample size in the examples provided; however, as will be demonstrated in the comparison with ABC methods, increasing the size of this sample directly improves the ability of the Fourier estimator to recover the true posterior, whereas because of the approximations involved in practice with ABC, such as summary statistics, it is unclear whether this increase in the value of N improves the estimation

insofar as recovering the true posterior is concerned. It is also possible for the Fourier approach to use summary statistics, although it would not additionally need to assume a multivariate normal distribution for them, again, reducing the amount of approximation.

Furthermore, the estimator itself has a different form from the one in [23], in keeping with the conditions established in the exact-approximate likelihood literature [4, 3], thus requiring additional theoretical and practical considerations. Here we also adopt the term "exact-approximate," as in [4, 3], in the sense that while the likelihood is approximated, in theory in can be done so arbitrarily accurately.

Using an estimator based on the Fourier integral theorem, we provide an exact simulation-based approach that extends to any model for which the likelihood is intractable, or otherwise unavailable, but from which simulation of data is possible. In contrast to the assortment of possible tuning parameters which must be chosen by information-reduction approaches, the tuning parameters in the proposed Fourier approach are simple, with well-defined behavior. By not relying on conversions to summary statistics or applications of a statistical distance function, we moreover maintain a greater sense of the reliability of the Fourier approach, as secured by the properties of the Fourier integral, in comparison with that of alternative methods, which involve potentially uncertain approximations and substantial information loss.

1.1. Fourier integral theorem

To begin, we sketch the Fourier approach in the ABC setting, introducing the concepts for which a more formal presentation will be given in Section 3.

To obtain an estimate of the likelihood, we can directly apply the Fourier integral theorem so that the posterior distribution is given exactly as

$$\pi(\theta|y) = \frac{\pi(\theta)}{(2\pi)^n\, m(y)} \int\int \cos(s\cdot(y-x)) f(x \mid \theta)\, ds\, dx,$$

where $m(y)$ is the marginal likelihood for observed data $y = (y_1, \ldots, y_n)$ and $\pi(\cdot)$ is the prior density. This result follows since the s integral can be done precisely on $(-R, R)^n$ and the x integral is done using Monte Carlo. That is, $\pi(\theta \mid y)$ can be estimated via Monte Carlo methods through an i.i.d. sample from $f(\cdot \mid \theta)$.

For some fixed value of R, which will be discussed later, let

$$f_R(y|\theta) = \int \prod_{i=1}^{n} \frac{\sin(R(y_i - x_i))}{\pi(y_i - x_i)}\, f(x|\theta)\, dx;$$

then, we define a "Fourier likelihood" function by

$$f_{R,N}(y|\theta) \propto \frac{1}{2}\left[f_R(y|\theta) + E\left| \frac{1}{N}\sum_{k=1}^{N}\prod_{i=1}^{n} \frac{\sin(R(y_i - x_{ik}))}{\pi(y_i - x_{ik})} \right| \right], \tag{1}$$

where the expectation is for the i.i.d. $\{x_{ik}\}$ of size $n \times N$ from $f(\cdot \mid \theta)$. The corresponding posterior density is thus given by $\pi_{R,N}(\theta|y) \propto \pi(\theta)\, f_{R,N}(y \mid \theta)$.

We take an unbiased Monte Carlo version of (1) as

$$\hat{f}_{R,N}(y \mid \theta) = \max\left\{0, \frac{1}{N}\sum_{k=1}^{N}\prod_{i=1}^{n}\frac{\sin(R(y_i - x_{ik}))}{\pi(y_i - x_{ik})}\right\}. \tag{2}$$

It is straightforward to demonstrate the unbiasedness based on $\max(0, a) = (a + |a|)/2$. This estimator is thus compatible with the "noisy likelihood" or "pseudo-marginal" approach [4], which guarantees the posterior based on (1) to be the correct stationary density when using (2) to stand in for (1). In Section 2, we provide background on this approach. Furthermore, as we reveal in Section 5.4, a KDE approach with a Gaussian kernel is unable to compete with the Fourier approach, especially as the dimension increases.

After Section 3 establishes properties of the estimator, Section 4 discusses how to choose the values of R and N to ensure that the estimation is highly accurate, because we can choose R and N such that the estimate is arbitrarily close to the true likelihood function.

So far, we have assumed y is univariate, but the extension to higher dimensions is straightforward; there, we would have the component for observation $y_i = (y_{i1}, \ldots, y_{id})$ as

$$\prod_{j=1}^{d}\frac{\sin(R(y_{ij} - x_{ikj}))}{\pi(y_{ij} - x_{ikj})}, \tag{3}$$

where d is the dimension of y. The important feature is that, based on the Fourier integral theorem, the estimator automatically picks up the dependence structure present in the samples, whereas a Gaussian kernel would typically require an estimated covariance matrix.

In general, the bias of the estimator can be of order $\exp(-R^2)$, and the variance is of order R^d/N (see Lemma 1). Given that we are simulating data from the model, we can take a reasonably large number of samples N so that a choice of R in the range of 5 to 20 works well, although the particular choice of R becomes less important as the size of N increases, especially after a normalizing transformation of the data, such as that described in Section 4.

Finally, we emphasize that the bias of order $\exp(-R^2)$ arises when the data have a mixture of normal appearance, and often it is sufficient to transform the data to have a normal appearance, assuming unimodal behavior, as will be demonstrated in Section 5.

1.2. Alternative approaches

We begin by considering a form of diagram which follows that of Figure 2 in [30]. First, in Figure 1, we show a diagram of the ABC likelihood approach, with notation mostly following that of [3]. In practice, ABC begins with the data being transformed to summary statistics, i.e., $s : \mathcal{Y} \to \mathbb{R}^q$, for some positive

integer q. Then, another function ψ_{d,K_ϵ} must be specified, requiring the choice of distance function $d(\cdot)$ and kernel $K_\epsilon : \mathbb{R}_{\geq 0} \to [0,1]$, such that $\psi(y, y') = K_\epsilon(d(s(y), s(y'))$, for $y, y' \in \mathcal{Y}$ and ϵ the bandwidth of the kernel. Next, in Figure 2, we show a diagram of the SL approach of [30]. Like ABC, SL depends on a transformation to summary statistics, s; however, SL further assumes that these statistics are multivariate normal; i.e., $s \sim \mathcal{N}(\mu_\theta, \Sigma_\theta)$, with the quantities μ_θ and Σ_θ also generally needing to be estimated. Finally, in Figure 3, we show the corresponding diagram of the Fourier approach.

All approaches start at the top, with the goal of evaluating the fit of the model with parameter vector θ to the raw data vector y. (Here, for direct comparison with SL, we adopt the notation used in [30], with the simulated data vectors y_i^* corresponding to the vectors x_i referred to in Section 1.1.) From the model, N data vectors $y_1^*, \ldots, y_N^*$ are simulated, given the value of θ. The Fourier approach then gives immediate access to the estimate of the log likelihood, using (2). In contrast, ABC and SL at this step first transform the data to summary statistics.

SL uses the statistics to estimate the mean vector and covariance matrix of s, according to the model with parameter vector θ, which are then supplied as arguments of the log multivariate normal (MVN) probability density function to evaluate the log synthetic likelihood. ABC applies the ψ function to the statistics based on the simulated and observed data to estimate the "ABC likelihood" function. The choice of summary statistics is crucial to SL and ABC, with the optimal choice generally being problem dependent. The requirement imposed by SL that these statistics be normal and the additional choices of kernel and distance functions required by ABC complicate matters further. Thus, we see that both ABC and SL involve a greater degree of approximation and tuning than does the Fourier approach. Further papers on the ABC approach include [22, 11, 25, 19], for example. These papers are principally concerned with summary statistic ideas. In high-dimensional settings, ABC can be implemented component-wise, as in a Gibbs framework, using conditional densities, as [9] demonstrate, although the theoretical justification differs from that of the standard Gibbs sampler. Such a strategy can also be done with the Fourier approach.

To avoid the use of summary statistics, full-data ABC approaches, as discussed by [10], omit the transform in Figure 1 and compare data sets on the basis of their distributions, using measures of distance such as the Wasserstein distance [7], the energy distance [20], the maximum mean discrepancy [21], the Cramer-von Mises distance [12], and the Kullback-Leibler divergence [15]. While these approaches avoid the potentially "extensive loss of information" incurred by the reduction of the data to summaries, the effectiveness of any choice of distance depends on the "size, type, and variability of the data," as [10] point out. Furthermore, despite using the full data, distance-based approaches can still result in information loss, especially with dependent data. To overcome this unreliability, one might consider introducing summary statistics or combining multiple distance functions in a weighted average, although clearly such modifications undermine the premise of the full-data distance-based approaches and can result in greater incertitude about the nature of the estimation.

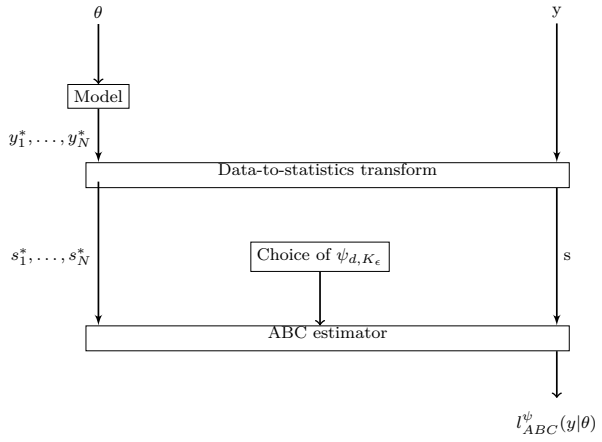

FIG 1. *Diagram of ABC.*

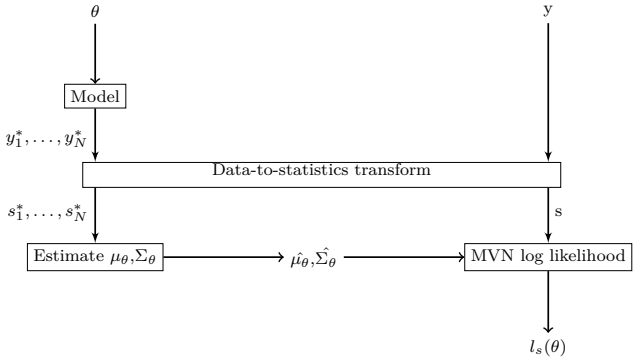

FIG 2. *Diagram of synthetic likelihood.*

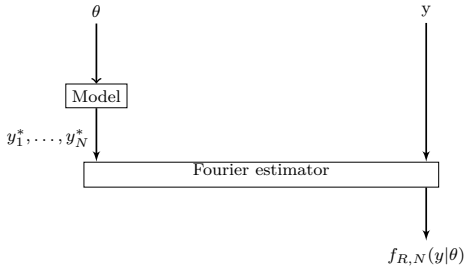

FIG 3. *Diagram of Fourier approach.*

The remainder of the paper is as follows. Section 2 provides background on MCMC estimation using the Fourier approach. Section 3 presents theoretical results. Section 4 covers implementation details and, in particular, focuses on the setting of the parameters R and N. Section 5 contains a number of examples. Section 6 features a brief conclusion. An Appendix includes mathematical results from Section 3, as well as additional results for Sections 5.1 and 5.3.

2. Background on "noisy" MCMC algorithms

This section considers the implementation of an MCMC algorithm, specifically a Metropolis sampler, when the likelihood is represented by a Monte Carlo approximation, a "noisy" likelihood. As noted by [17], for models that lack a likelihood function, for either mathematical or computational reasons, Bayesian approaches must resort to "practical if cruder approximation methods," such as Laplace approximations [29] and variational Bayes solutions [14]. [17] then tout ABC as offering an "almost automated resolution" to this class of models, while acknowledging that "these methods suffer to some degree from calibration difficulties that make them rather volatile in their implementation and thus render them suspicious to the users of more traditional Monte Carlo methods." That is, despite the careful selection and calibration of a number of underlying parameters, a process that in itself takes some guesswork, methods based on summary statistics, like SL, can exhibit unpredictable behavior for which a remedy is unclear. By not requiring a transformation to summary statistics and by relying on tuning parameters that relate directly to the Fourier integral theorem, the Fourier approach provides a clear advantage over such methods, both conceptually and practically.

We begin by introducing the main problem faced by Bayesian methods in attempting to implement MCMC algorithms, of which we take the Metropolis–Hastings algorithm [13] as the primary representative, when the likelihood function cannot be evaluated. To apply the Metropolis–Hastings algorithm, we need to evaluate the acceptance ratio, the general form of which, for parameters θ and data y, is given by

$$\alpha(\theta, \theta') = \frac{\pi(\theta'|y)h(\theta|\theta')}{\pi(\theta|y)h(\theta'|\theta)},$$

where π is the desired target distribution and h a proposal distribution. When the form of the likelihood is either unknown or unavailable, we can turn to an approximation of α.

Described in [5, 4], the "pseudo-marginal" approach derives an approximation of α by estimating the target distribution, π, at each θ. With this approach, given θ' drawn from $h(\cdot|\theta)$, followed by x_i drawn i.i.d. from the likelihood function $f(\cdot|\theta')$, $i = 1, \ldots, N$, the approximate acceptance ratio $\widehat{\alpha}$ takes the form

$$\widehat{\alpha}(\theta, \theta') = \frac{\widehat{\pi}(\theta'|y)h(\theta|\theta')}{\widehat{\pi}(\theta|y)h(\theta'|\theta)},$$

where $\widehat{\pi}$ represents the prior times an approximation of the likelihood.

[4] show that once this procedure defines an irreducible and aperiodic Markov chain, it will produce samples distributed asymptotically according to π. Moreover, the same result applies to any (non-negative) unbiased estimator of π.

Two schemes for estimating the target distribution are discussed by [4]: Monte Carlo within Metropolis (MCWM) and grouped independence MH (GIMH). At each iteration, MCWM computes the estimate of the target distribution for both the current θ and the proposed θ'. GIMH retains the estimate from the previous iteration and computes only the estimate for the proposed parameter value. The difference between these schemes is that GIMH produces samples distributed asymptotically according to the desired target distribution, whereas MCWM produces samples distributed asymptotically according to an approximation of the target distribution.

Following the analysis of ABC-MCMC in [3], we consider, for the Fourier case, the nonnegative and unbiased estimator given by (2). With the standard "auxiliary variable trick" identified by [3], the Fourier approach, using this estimator, is then exact in the twofold sense that for an ergodic Markov chain based on GIMH, samples can be produced so as to be arbitrarily close to the approximated posterior distribution, with convergence to the true posterior as $R \to \infty$, assuming nonnegative estimates. In particular, approximating the likelihood function with the Monte Carlo estimator (2), we regard our stationary density as proportional to the prior times the approximation to the likelihood, i.e., (1). What gives the Fourier approach a significant advantage is that it can pick up dependence structures without an estimated covariance matrix, allowing for highly accurate estimates in higher dimensions.

While the Fourier approach can be implemented with either MCWM or GIMH, MCWM can help the chain to avoid getting stuck for any extended period, thereby allowing for greater efficiency than GIMH. Indeed, as [2] state, "[D]espite its exactness, there is no particular reason for estimators from GIMH to be more statistically efficient than those from MCWM."

The implementation of any algorithm for estimating a likelihood using the Fourier estimator can be further improved by adopting the sampling techniques suggested by [3] to prevent the aforementioned "sticky behavior." The concern here shifts from bias to variance, where performance improvements are obtained by reducing the variability of the posterior estimates in terms of, say, the convex order, which [3] maintain as a natural quantity in this context, backed by both intuition and theory.

One such strategy is to average the estimators for each proposed value of θ in the MCMC, a procedure that can be done in parallel at little extra cost; that is, the estimation of the likelihood is performed M times according to the evaluation schemes in Section 1.2, after which the estimates are averaged. For models for which sampling time can increase steeply as a function of N, this strategy can help to reduce the number of samples needed in each parallel instantiation.

In the next section, we present theoretical results that will illuminate ways of dealing with observed points of low density, such as points in the tails of a distri-

bution, and with smaller simulated sample sizes. In contrast with the marginal likelihood evaluation predominant in [23], in which one may specify any high-density point, likelihood evaluation will likely involve such low-density points, especially as the sample size increases. For computational efficiency, because the evaluation must be performed at each iteration of the MCMC, the sample size of simulated data on which the Fourier estimation is based also tends to be smaller for these problems than for those in [23]. As previously mentioned, we will also be working with an estimator whose form is different from the one in [23].

3. Theory

In this section, we begin with some theoretical results, expanding on the introductory remarks in Section 1.1. As will be made apparent in the following presentation, there comes a fundamental shift in the focus of the Fourier approach, from bias in the marginal likelihood setting to efficiency in the ABC setting. Proving exceedingly helpful in establishing guidelines for setting the values of R and N in (2), these results, based on the assumption of normal data, also apply to mixtures of normals, and so cover quite an extensive class of model; moreover, in Section 4.2, we introduce a procedure to effect the normal transformation, based on the Yeo-Johnson transformation [31], a power transformation that is shown in the examples in Section 5 to work well in practice. Bear in mind, however, that to apply the Fourier approach, we do not require that the data be normal.

Now, let us recall the results taken from [23].

Definition 1. *For some fixed value of R, define the Fourier integral representation of $f(y_1, \ldots, y_n \mid \theta)$ as*

$$f_R(y|\theta) = \int \prod_{i=1}^{n} \frac{\sin(R(y_i - x_i))}{\pi(y_i - x_i)} f(x|\theta)\, dx. \tag{4}$$

This becomes more precise as $R \to \infty$.

Definition 2. *As in [23], the Fourier estimator of (4), based on a Monte Carlo sample $(x_k = (x_{1k}, \ldots, x_{nk}))$ from $f(\cdot \mid \theta)$, with $n = 1$, is given by*

$$\tilde{f}_{R,N}(y) = \frac{1}{N} \sum_{k=1}^{N} \frac{\sin(R(y - x_k))}{\pi(y - x_k)}. \tag{5}$$

For this estimator (5), we observe the following asymptotic behavior.

Lemma 1 ([23])**.**

$$\mathit{Bias}\left(\tilde{f}_{R,N}(y)\right) = O\left(\exp\left(-\sigma^2 R^2/2\right)\right) \quad \text{and} \quad \mathit{Var}\left(\tilde{f}_{R,N}(y)\right) = O(R/N).$$

While this behavior is naturally relevant for larger values of R, with appropriately large values of N, further insight into the dependence of the estimator

on these parameters is afforded by the following theoretical results, which give precise formulations for understanding the behavior of the estimator for any values of R and N; that is, with these formulations, we can calculate the exact values for the mean and variance of the estimator, assuming normal data.

These formulations are based on the following two equations: For $X \sim \mathcal{N}(\mu, \sigma^2)$, let

$$I_1 \equiv \frac{\pi}{2}\mathcal{N}(y|\mu,\sigma^2)\left[\operatorname{erf}\left(\frac{\sigma R - \frac{i(y-\mu)}{\sigma}}{\sqrt{2}}\right) + \operatorname{erf}\left(\frac{\sigma R + \frac{i(y-\mu)}{\sigma}}{\sqrt{2}}\right)\right], \tag{6}$$

and

$$I_2 \equiv \frac{\pi}{2}\mathcal{N}(y|\mu,\sigma^2)(A_1 + A_2 + A_3) - C_1, \tag{7}$$

where

$$A_1 \equiv \frac{\exp(-2\sigma^2[R - \frac{i}{2\sigma^2}(y-\mu)]^2) + \exp(-2\sigma^2[R + \frac{i}{2\sigma^2}(y-\mu)]^2)}{\sqrt{2\pi\sigma^2}};$$

$$A_2 \equiv [R + \frac{i}{2\sigma^2}(y-\mu)]\operatorname{erf}\left(\frac{R + \frac{i}{2\sigma^2}(y-\mu)}{\sqrt{\frac{1}{2\sigma^2}}}\right);$$

$$A_3 \equiv [R - \frac{i}{2\sigma^2}(y-\mu)]\operatorname{erf}\left(\frac{R - \frac{i}{2\sigma^2}(y-\mu)}{\sqrt{\frac{1}{2\sigma^2}}}\right);$$

$$C_1 \equiv \frac{1}{2\sigma^2} + \operatorname{Re}\left(i\frac{\sqrt{\pi}}{2\sigma^2\sqrt{2\sigma^2}}(y-\mu)\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)\operatorname{erf}\left(i\sqrt{\frac{1}{2\sigma^2}}(y-\mu)\right)\right).$$

Accordingly, we now present the main theoretical results of the paper.

Theorem 1. *The first and second moments of $\tilde{f}_{R,N}(y)$ are given by (6) and (7), respectively.*

Because this result is uniform for the normal distribution with mean μ and variance σ^2, it also holds for a mixture of normal distributions.

As with rules-of-thumb for setting bandwidth parameters in kernel density estimation, such as that due to Silverman [28], based on normal approximations, we next derive a relationship between N and R which provides guidance on setting the value of R given a value of N, guidance motivated by the computational concerns that encourage smaller values of N, to be set first with a view to efficient computation, followed by the optimal choice of R to reduce the MSE. In Sections 4.1–4.2, we use these theoretical results to examine the relationship between R and N more fully and conduct numerical studies to substantiate the proposed rule-of-thumb, in addition to demonstrating normalizing transformations as part of an automatic procedure for applying the Fourier approach.

Theorem 2. *With the standard normal distribution, as $N \to \infty$, we have that $MSE\left(\tilde{f}_{R,N}(0)\right)$ is minimized when $erf\left(R/\sqrt{2}\right) = 1$.*

Corollary 1. *The optimal selection of R is for $R \approx 5.923 \cdot \sqrt{2}$.*

The advantage of the Fourier approach is the relatively low values of R necessary for obtaining an optimal result, with respect to the MSE. The low bias of the estimator for even moderate values of R, such as $R = 4$, allows for the variance to be controlled even for the smaller values of N, such as $N = 10^4$, which support the computational efficiency of the MCMC algorithm. Yet it should be stressed that both R and N come with clear guidelines and that increasing N, and thus the optimal choice of R, will directly improve the ability of the Fourier approach to recover the true posterior density in the ABC setting, an outcome that, as will be shown in Sections 5.2 and 5.4, relating to univariate and multivariate mixture models, respectively, cannot be claimed for other methods.

From here, we observe the aforementioned shift in focus from the Fourier estimator of [23] to the estimator of the "Fourier likelihood" function, which can attain more favorable MSE, a priority since with an MCMC sampler, an estimate of the likelihood is needed at each iteration of the sampler. Indeed, while the efficiency of the Fourier estimator is generally a convenience for marginal likelihood estimation, it is nearly a necessity for intractable likelihood estimation in a "noisy" MCMC framework, as introduced in Section 2.

Recall that this framework requires a non-negative, unbiased estimator of the target posterior. To derive this posterior, we thus introduce a new likelihood function, the "Fourier likelihood" function, along with an unbiased estimator of this function.

Definition 3. *The "Fourier likelihood" function is given by*

$$f_{R,N}(y|\theta) \propto \frac{1}{2}\left[f_R(y|\theta) + E\left|\frac{1}{N}\sum_{k=1}^{N}\prod_{i=1}^{n}\frac{\sin(R(y_i - x_{ik}))}{\pi(y_i - x_{ik})}\right|\right], \tag{8}$$

where the expectation is for the i.i.d. $\{x_{ik}\}$ of size $n \times N$ from $f(\cdot \mid \theta)$.

Corollary 2. *The Monte Carlo estimator*

$$\hat{f}_{R,N}(y \mid \theta) = \max\left\{0, \frac{1}{N}\sum_{k=1}^{N}\prod_{i=1}^{n}\frac{\sin(R(y_i - x_{ik}))}{\pi(y_i - x_{ik})}\right\} \tag{9}$$

is an unbiased Monte Carlo version of (8).

Corollary 3. *The corresponding posterior density is given by $\pi_{R,N}(\theta|y) \propto \pi(\theta)$ $f_{R,N}(y \mid \theta)$.*

Accordingly, in the "noisy" MCMC framework, as discussed in Section 2, we can estimate this target posterior distribution using MCWM or GIMH.

The proofs of the results appear in the Appendix.

Let us now proceed to examine some implications of these results for intractable likelihood estimation. According to Theorem 2, we can optimally choose $R \approx 5.923 \cdot \sqrt{2}$; this theorem serves as the basis for our setting the range of R values between 5 and 15 for data that have a normal appearance.

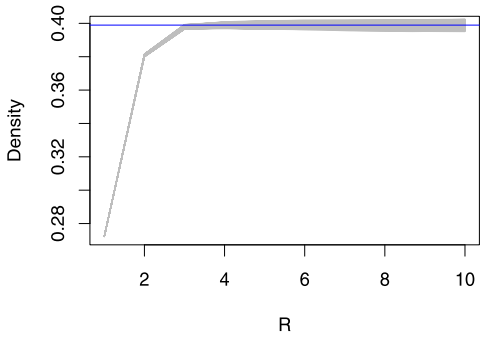

FIG 4. *Estimates of a standard normal density at the value 0.0 over* 10^4 *independent simulations. The blue line is the true value.*

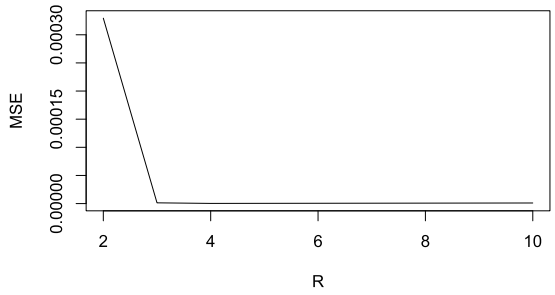

FIG 5. *Estimates of the MSE.*

Nevertheless, we are not claiming that the data need to have this appearance for the Fourier approach to be applicable, but only that for the automatic procedure that we have here adopted, this assumption helps to ensure the selection of a reasonable range of R values.

Using Theorem 2, we further find that, with, say, $N = 10^6$, MSE $\left(\tilde{f}_{R,N}(0)\right)$ is minimized for $R \approx 3.6$. In Figure 4, we plot the Fourier estimates of the standard normal density at $y = 0$ over 10^4 independent simulations as a function of R, evaluated at integer values from 1 to 10, and in Figure 5, we plot the corresponding estimates of the MSE values. The minimum value occurs, on the aforementioned integer grid, at $R = 4$.

As previously stated, in the ABC implementation, we are likely to encounter low-density points and work with smaller sample sizes N. Yet, with the estimator in (9), the resulting MSE, with respect to the true density, can be improved over that of the estimator in (5). To illustrate, we take $N = 10^3$ and the observed value to be $y = 3.0$. In Figure 6, we plot the Fourier estimates from the same simulation as above but with the new specification of N and y. We observe that the estimates due to (5) display a greater variability than do those due to (9) as R increases, with a higher incidence of negative values for larger values of R. The bias, however, remains controlled in both cases. Consequently, the estima-

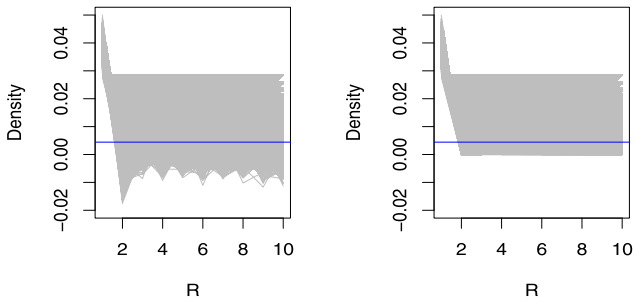

FIG 6. *Estimates of a standard normal density at the value 3.0 over* 10^4 *independent simulations. The blue line is the true value. Left: Estimates using* (5). *Right: Estimates using* (9).

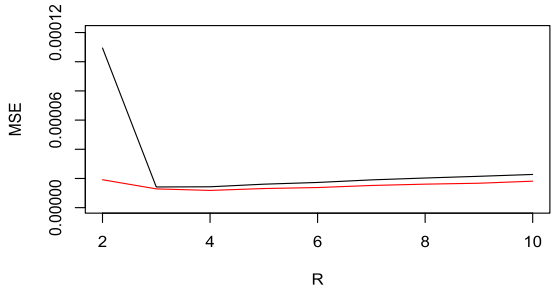

FIG 7. *Estimates of the MSE. The black lines represents* (5) *and the red,* (9).

tor in (9) would appear to be more viable with smaller values of R, say, R in the range from 2 to 4, than the estimator in (5). To corroborate these observations, in Figure 7, we plot the corresponding estimates of the MSE values. For the estimator in (5), we find that the MSE is minimized at $R = 3$ and for the estimator in (9), $R = 4$; moreover, the estimate of the MSE is lower in the case of (9).

To conclude, we note the following inequality, which, in establishing the connection between the estimators (5) and (9), serves to encapsulate the preceding relationships and guide the implementation of the Fourier approach. Given that $|x| = 2\max\{0, x\} - x$, for $n = 1$, we see that

$$\begin{aligned}
\left|E\left[\hat{f}_{R,N}(y|\theta)\right] - f_R(y|\theta)\right| &= \left|\frac{1}{2}\left[f_R(y|\theta) + E\left|\tilde{f}_{R,N}\right|\right] - f_R(y|\theta)\right| \\
&= \frac{1}{2}\left|E\left|\tilde{f}_{R,N}\right| - f_R(y|\theta)\right| \\
&\leq \frac{1}{2}\left|\sqrt{\mathrm{Var}\left(\tilde{f}_{R,N}\right) + f_R(y|\theta)^2} - f_R(y|\theta)\right| \\
&\leq \frac{1}{2}\left|\sqrt{\mathrm{Var}\left(\tilde{f}_{R,N}\right) + f_R(y|\theta)^2} - f(y|\theta)\right| \\
&\quad + \frac{1}{2}\left|f(y|\theta) - f_R(y|\theta)\right|, \qquad (10)
\end{aligned}$$

by Jensen's inequality and the triangle inequality. Recognize that $f_R(y|\theta) \to f(y|\theta)$ as $R \to \infty$, so that by keeping the variance of $\tilde{f}_{R,N}$ low, we can improve the approximation to the true likelihood.

Recall from Section 2 that one strategy to reduce the variability of a "noisy" MCMC implementation is to average the estimator over M independent instantiations at each iteration of the sampler. In the Fourier case, we may thus choose to average each such estimator $\tilde{f}^{(m)}_{R,N}$ (as in (5)) such that

$$M^{-1}\sum_{m=1}^{M}\max\left\{0,\tilde{f}^{(m)}_{R,N}\right\} \quad \text{or} \quad \max\left\{0,M^{-1}\sum_{m=1}^{M}\tilde{f}^{(m)}_{R,N}\right\}.$$

In the implementation details in the following section, we provide techniques to attain estimates that have a high probability of being positive, even for low-density points and small replicated sample sizes. When coupled with such techniques, the latter choice of estimator is preferable inasmuch as it reduces the variance referred to in (10); therefore, it is this variation that we use in the examples in Section 5.

4. Implementation

In this section, we further discuss the implications of the theoretical results. First, in Section 4.1, we provide an analysis of the two tuning parameters, R and N, underlying the Fourier approach. Then, in Section 4.2, we discuss the transformation to normal-looking data, on which the guidance in Section 4.1 is predicated. Finally, in Section 4.3, we give an assessment of conditional decompositions that facilitate the modeling of higher-dimensional data.

4.1. Choosing the values of R and N

In choosing the value of R in the Fourier likelihood estimator given by (9), we can first apply a normalizing transformation to the relevant data, which can help to standardize the choice of R values. We begin with a description of a procedure using a Yeo-Johnson transformation [31], a power transformation defined on the whole real line, to evaluate the density of the standard normal distribution at the point $y = -3.008$, corresponding to the minimum value obtained from a random sample in R, as generated from the function *rnorm(1e4)* with *set.seed(1)*: To these data, we apply a Yeo-Johnson transformation based on the data falling below the 0.05 empirical quantile; then, we center and scale, so that the appropriate R values can be more readily identified. (See `https://github.com/FARotiroti/Fourier`: this whole process can be done automatically, with a certain quantile specified.) Note that the proposed transformation is to help to standardize the choice of R value; it is not a necessary part of the Fourier approach. As we will see with the Ricker model in Section 5.5, even for data that are non-normal after transformation, the approach is still able to produce accurate estimates.

In general, because of the standardization effected by this procedure, we take the median of the estimates over the range $R = 5$ to 15, accounting for the behavior of the estimator shown in Figure 2 of [23], whereby it is revealed that once the estimates have stabilized over a particular region of R values, the particular value of R chosen within that region has little effect on the estimate. Given the automatic procedure we are here advocating, using a normalizing transformation, followed by centering and scaling operations, the range from $R = 5$ to 15 generally encompasses a stable range, and the median adds further robustness to the estimator, for either the estimates have stabilized sufficiently or they are distributed in a sinsoludal pattern around the true value; in either case, the median helps to capture the desired estimate.

While we could perform the estimation repeatedly and then average, as discussed at the end of Section 2, a single estimate is generally sufficient in an MCMC implementation given that this value contributes very little to the overall likelihood of the 10^4 observed data points, yet for illustrative purposes, we are here interested in obtaining a precise value.

The variance of the estimate at the value of -3.008 when $R = 10$ and $N = 10^4$ is 2.24e-06, with mean 0.004. Ideally, one would search for the relevant partition such that after transformation each point could be treated as roughly normal, although practically this strategy is not feasible, especially for the sampling method used below. Moreover, while we might use the full data as a matter of convenience, here we sample 10^4 data points only below the 0.05 quantile. After transformation, the minimum value becomes -1.931. Assuming normality, the density estimate of this value, with 10^4 data points, at $R = 10$, has mean 0.062, with variance 2.1e-05. Over 10^5 independent simulations, the mean of the estimate using transformed data is 0.061, with variance 2.1e-05. Since the Jacobian cannot cause a negative value, we conclude that the estimate is very unlikely to be negative here. Thus, we have produced an estimate effectively bounded away from 0.

The boxplot in Figure 8 shows a comparison between estimates based on applying the Fourier approach with and without transformation, when $R = 10$ and $N = 10^4$, over 10^5 independent simulations. The black line is the true density. With the original approach, only 0.13% of the estimates are negative, so as expected for normal data, even for estimation in the tails, the direct application of the Fourier estimator produces reliable estimates on average; still, the transformed approach can be seen to yield estimates with lower variance, which are never negative in this example.

Consider that for a given value of R, the true Fourier integral can be negative for values in the tails. Observe the expected values of the estimator for different values of R in Figure 9. Thus, the first task is to ensure that we have a sufficiently large R value. Assuming a finite data set drawn i.i.d. from a standard normal, we can use the results from Section 3 to ensure that such an R value is chosen for each observed value in the data set.

Here, let us focus on a tail value like $y = 3.5$. Continuing to use the numerical approximation to the erf function due to [1], we determine that the values of the integral are -0.0003 for $R = 1$, 0.0002 for $R = 2$, and 0.0006 for $R = 3$. The

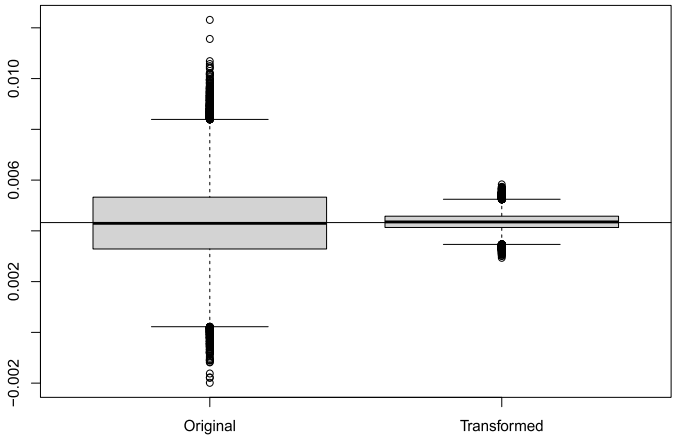

FIG 8. *Estimates of a standard normal density at the tail value of* -3.008 *over* 10^5 *independent simulations, using the Fourier approach on the original data and on transformed data, both with* $R = 10$ *and* $N = 10^4$.

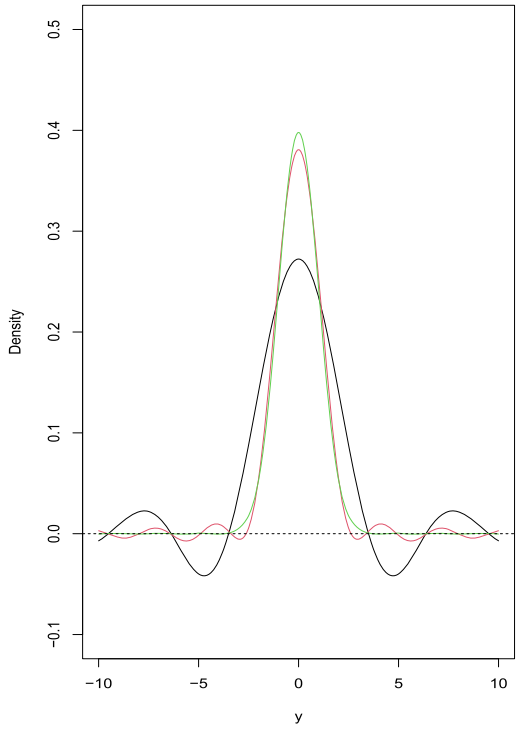

FIG 9. *Fourier integrals. The black line represents the values for* $R = 1$, *red* $R = 2$, *and green* $R = 3$.

values of R need not be integers, but we are using integers here for illustration. Also note the previously mentioned sinusoidal pattern that emerges in the tails.

In practice, we will not know which R value is sufficiently large. So, in this instance, suppose we take $R = 5$, since for normal-looking data, this value is reasonable for estimating nearly any likely observation. For $R = 5$, we, however, know that the expected value is 0.0009, which is the true value.

Having chosen our R value, we can then consider the variance of the estimator to guide our choice of N value. For $N = 1$, the variance is 0.008, indicating that the estimator is likely to produce negative values. So suppose that we stipulate that the estimator remain positive within two standard deviations of its expected value. For $R = 5$, we find that this criterion is met around $N = 50000$. If we increased the stipulation to three standard deviations, $N = 10^5$ is an appropriate value. Thus, supposing that $y = 3.5$ is the most extreme in the observed data set, we could comfortably set $R = 5$ and $N = 10^5$ to achieve an unbiased estimator of the true density value which is seldom negative.

To substantiate this analysis with numerical results, we estimated the density at $y = 3.5$ over 10^6 replications, each with $R = 5$ and $N = 10^5$. Over these replications, the mean was 0.0009, with standard deviation 0.0003, as compared with the expected mean of 0.0009, with standard deviation 0.0003. The estimator was negative 0.063% of the time.

While the preceding examples concerns the estimation of the density in the tail, the normalizing transformation can be applied to improve the estimation of the density at any point, this automatic procedure requiring only the specification of a partition of the data. Taking the median of the resultant estimates over the range of R values between 5 and 15 further delivers robust estimation.

4.2. Yeo-Johnson transformations

While normality of the data is not a requirement to apply the Fourier approach, the theory developed in Section 3 helps to guide its implementation when the data appear normal. As illustrated in the previous section, to effect an automatic method, based on that theory, we propose the application of a Yeo-Johnson transformation to the data, as described in [31]. Such a transformation is part of a power transformation family aimed at achieving approximate normality and reducing skewness. It is defined as $\xi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that

$$\xi(\lambda, x) = \begin{cases} \{(x+1)^{\lambda} - 1\}/\lambda, & (x \geq 0, \lambda \neq 0), \\ \log(x+1), & (x \geq 0, \lambda = 0), \\ -\{(-x+1)^{2-\lambda} - 1\}/(2-\lambda), & (x < 0, \lambda \neq 2), \\ -\log(-x+1), & (x < 0, \lambda = 2). \end{cases}$$

We have chosen to use this transformation because it is well defined on the entire real line, giving it general applicability. The choice of λ follows the procedure outlined in [31], which depends on maximum likelihood estimation under the assumption that for some λ, the transformed observations can be modeled as following a normal distribution.

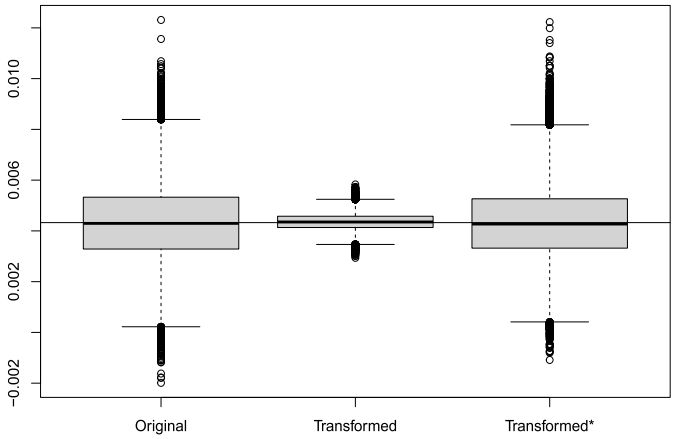

FIG 10. *Estimates of a standard normal density at the tail value of -3.008 over* 10^5 *independent simulations, using the Fourier approach on the original data and on transformed data, both with* $R = 10$ *and* $N = 10^4$.

Resuming the discussion in Section 4.1, we consider the values that underlie Figure 8. We obtain $\lambda = 4.480$, obtained by considering only the data that fall below the empirical 0.05 quantile of the data, drawn i.i.d. from a standard normal with size 10^3, namely, below -1.727, given data generated from *set.seed(1)* in R. The estimates are standardized using the mean and standard deviation of this subset, -0.380 and 0.006, respectively. Thus, with reference to the third case of the function ξ disclosed above, the absolute value of the Jacobian is given by $|(-x+1)^{1-\lambda}|$ scaled by the aforementioned standard deviation. This quantity must then be multiplied by the proportion of the data used.

In the examples in the paper, we do not use just this proportion of the data but rather the entire data set for each estimate. If, say, we wanted to use 10^4 samples to perform the estimate, we would need to estimate $10^4/0.05$ values. In Figure 10, we add to Figure 8 an additional boxplot (labeled Transformed*) which simply uses the full data size of 10^4, so that the size of the sampled data is the same for Original and Transformed*. Although the estimates are very similar between those two approaches, for Transformed*, the estimates are negative 0.04% of the time, whereas for Original, the estimates are negative 0.13% of the time. Still, keep in mind that the data here are normal, and the greatest gains will be realized when the data must be transformed, such as with the Ricker data in Section 5.5, which are positively skewed count data.

4.3. Conditional decompositions

In Section 5, the Fourier estimator will be seen to perform well in higher dimensions, in comparison with alternative approaches, such as KDE, SL, and full-data distance-based methods, like CvM. In fact, in Section 5.5, we will con-

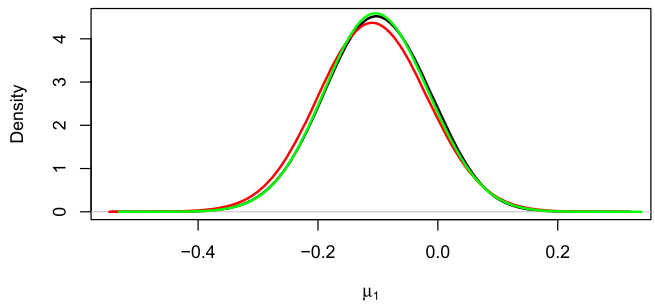

FIG 11. *Comparison of marginal posterior densities of μ_1 using the exact likelihood (black), a Fourier likelihood without blocking (red), and a Fourier likelihood with two blocks (green).*

sider the estimation of eight- and nine-dimensional density functions, rolling over 100 time points. Still, if sampling of constituent conditional distributions is available, the Fourier approach can be extended to problems of increasingly higher dimension.

As an example, suppose that $y \sim N(\mu, \Sigma)$, with $\mu = 0$ and

$$\Sigma = \begin{bmatrix} 0.97 & 0.47 & -0.51 & -0.04 \\ 0.47 & 1.89 & -1.58 & 0.07 \\ -0.51 & -1.58 & 1.60 & -0.01 \\ -0.04 & 0.07 & -0.01 & 0.48 \end{bmatrix}.$$

We draw a data set of 100 independent values from this distribution with $\mu = 0$. Then, assuming a flat prior on μ, we draw posterior samples in three ways:

1. Using the true likelihood;
2. Using the full four-dimensional vector;
3. Using two two-dimensional vectors formed from the first and last pairs of $y = (y_1, y_2)$ with the distributions given by y_1 and $[y_2|y_1]$.

In Figs. 11–14, the marginal posterior densities appear consistent across all approaches, with the approach with two blocks displaying a closer match with that using the true likelihood than does the approach without blocking. Nevertheless, the performance on the full four-dimensional vector suggests that the approach could thus be extended to, say, a 12-dimensional vector decomposed into three blocks, each consisting of a four-dimensional vector.

There are, moreover, other approaches to sampling the conditional distributions, with the central idea consisting in the evaluation of a density at a point. For instance, one could obtain samples from the posterior $[\theta|y_1]$ and then estimate the posterior density at the proposed draw in the MCMC algorithm using the Fourier estimator. If it is then assumed that the conditional distribution $[y_2|\theta, y_1]$ is known, this approach gives access to a highly useful means of obtaining samples from $[\theta|y_1, y_2]$; if it is not assumed known, then the Fourier estimator can again be used to estimate this conditional likelihood.

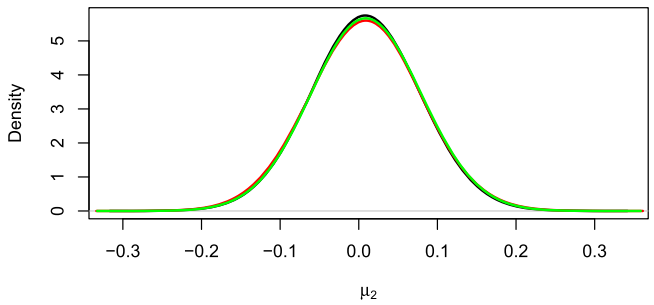

FIG 12. *Comparison of marginal posterior densities of μ_2 using the exact likelihood (black), a Fourier likelihood without blocking (red), and a Fourier likelihood with two blocks (green).*

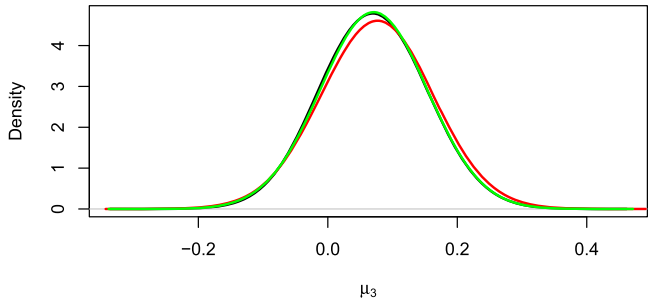

FIG 13. *Comparison of marginal posterior densities of μ_3 using the exact likelihood (black), a Fourier likelihood without blocking (red), and a Fourier likelihood with two blocks (green).*

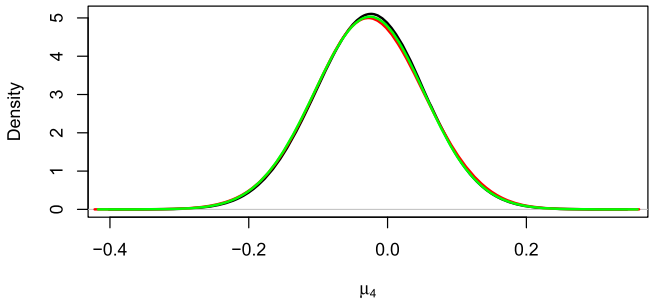

FIG 14. *Comparison of marginal posterior densities of μ_4 using the exact likelihood (black), a Fourier likelihood without blocking (red), and a Fourier likelihood with two blocks (green).*

5. Examples

We begin with two examples, the first involving the g-and-k distribution and the second an M/G/1 queuing model, both of which appear in [10]. We also consider an example involving the Ricker model, as described in [30].

For the examples that appear in [10], we will compare with only the best performing models there identified. See that paper for more details about the alternative approaches. Following both examples, we illustrate with mixture models common in the ABC literature how the Fourier approach is better able than the alternative approaches to represent the true likelihood. For the Ricker model example, we compare with synthetic likelihood (SL).

As we will demonstrate, the Fourier approach performs well across the various examples, using the full data without incurring information loss, while also remaining computationally viable. We will see that as opposed to the alternative approaches, the Fourier approach is properly estimating the true likelihood function, an outcome due to its solid statistical basis, which also aids in setting the couple of tuning parameters on which it relies. The data and code for all the examples is available at https://github.com/FARotiroti/Fourier.

Note that because of the differences in procedure between the Fourier approach and ABC methods, making an accurate comparison of the computation time is a difficult task. The comparison of such between the Fourier approach and both KDE and SL is more straightforward, however. In Section 5.3, for the M/G/1 queuing model, an estimate of the log likelihood is obtained on the average by the Fourier approach in 0.0072 seconds and by KDE in 0.0007 seconds. In Section 5.5, for the Ricker model, an estimate of the log likelihood is obtained on the average by the Fourier approach in 0.8775 seconds and by SL in 0.3256 seconds. In both these examples, the number of replicated data sets is $N = 10^4$.

It is thus evident that the methods considered are computationally viable in these examples; however, we emphasize that what sets the Fourier approach apart from these methods is its ability to estimate accurately the true posterior—with the tuning parameters playing clear roles in supporting the estimation—and to continue to perform well in higher dimensions.

5.1. g-and-k distribution

First, we consider the g-and-k distribution, which is defined in terms of its quantile function, as given by

$$Q(z(p)|\theta) = a + b\left(1 + c\,\frac{1-\exp(-g\ z(p))}{1+\exp(-g\ z(p))}\right) \times \left(1+z(p)^2\right)^k z(p),$$

where p denotes the given quantile and z the quantile function of the standard normal. The true parameter vector is set as

$$\theta = (a = 3, b = 1, g = 2, k = 0.05),$$

with $c = 0.8$ fixed, and the prior is taken to be $\mathcal{U}(0, 10)^4$. The heavy-tailed nature of the data makes it potentially challenging for certain full-data distance-based ABC approaches, like that using the Wasserstein distance, as well as for standard

TABLE 1
Simulation results, based on 50 data sets, for parameter a of the g-and-k distribution example with simulated data of size $n = 100$.

Method	mean	median	std	80%	90%	95%
Exact	3.018	3.008	0.119	0.88	0.92	0.96
Fourier	3.015	3.007	0.119	0.88	0.92	0.96
CvM	3.008	3.003	0.092	0.80	0.90	0.94

TABLE 2
Simulation results, based on 50 data sets, for parameter b of the g-and-k distribution example with simulated data of size $n = 100$.

Method	mean	median	std	80%	90%	95%
Exact	1.079	1.054	0.244	0.86	0.94	0.96
Fourier	1.077	1.055	0.241	0.86	0.96	0.98
CvM	1.048	1.034	0.195	0.76	0.92	0.96

TABLE 3
Simulation results, based on 50 data sets, for parameter g of the g-and-k distribution example with simulated data of size $n = 100$.

Method	mean	median	std	80%	90%	95%
Exact	2.252	2.143	0.600	0.88	0.98	0.98
Fourier	2.285	2.168	0.637	0.86	0.96	0.98
CvM	2.318	2.201	0.635	0.86	0.98	0.98

ABC. Because the likelihood can be computed numerically, we can compare the estimates with exact Bayesian results.

We consider sample sizes of $n = 100$ and $n = 1000$. In each case, we repeat the experiment over 50 simulated data sets, using the same data as [10]. According to [10], the full-data distance-based ABC approach using the Cramer-von Mises (CvM) distance performs very well across all the comparisons.

For the four parameters, respectively, Tables 1–4 contain results for $n = 100$, and Tables 5–8, for $n = 1000$. Based on 50 simulated data sets, the results include the mean of the posterior means, the mean of the posterior medians, the mean of the posterior standard deviations, and the coverage rates for nominal rates of 80%, 90% and 95%.

From these tables, it is clear that the Fourier approach provides estimates closely matched to the exact results. For the Fourier approach, to set the parameters for the Yeo-Johnson transformation, we partition the data based on empirical quartiles. Results for the estimated posterior correlations can be found in the Appendix; there, in Tables 13 and 14, we find further evidence that the Fourier approach is estimating the true posterior distribution.

TABLE 4
Simulation results, based on 50 data sets, for parameter k of the g-and-k distribution example with simulated data of size n = 100.

Method	mean	median	std	80%	90%	95%
Exact	0.505	0.495	0.136	0.82	0.90	0.98
Fourier	0.500	0.489	0.136	0.84	0.92	0.98
CvM	0.530	0.517	0.174	0.86	0.96	0.96

TABLE 5
Simulation results, based on 50 data sets, for parameter a of the g-and-k distribution example with simulated data of size n = 1000.

Method	mean	median	std	80%	90%	95%
Exact	2.996	2.995	0.036	0.80	0.94	0.96
Fourier	2.997	2.996	0.038	0.82	0.96	0.96
CvM	2.995	2.994	0.038	0.84	0.94	0.96

TABLE 6
Simulation results, based on 50 data sets, for parameter b of the g-and-k distribution example with simulated data of size n = 1000.

Method	mean	median	std	80%	90%	95%
Exact	0.999	0.996	0.0715	0.84	0.94	0.98
Fourier	1.001	0.998	0.075	0.88	0.96	0.98
CvM	0.998	0.996	0.078	0.88	0.98	0.98

TABLE 7
Simulation results, based on 50 data sets, for parameter g of the g-and-k distribution example with simulated data of size n = 1000.

Method	mean	median	std	80%	90%	95%
Exact	2.001	1.998	0.102	0.82	0.86	0.90
Fourier	1.999	1.997	0.107	0.84	0.88	0.94
CvM	2.021	2.016	0.143	0.92	0.98	0.98

TABLE 8
Simulation results, based on 50 data sets, for parameter k of the g-and-k distribution example with simulated data of size n = 1000.

Method	mean	median	std	80%	90%	95%
Exact	0.501	0.500	0.0408	0.82	0.90	0.98
Fourier	0.499	0.498	0.043	0.86	0.90	0.98
CvM	0.508	0.506	0.063	0.96	1.00	1.00

In the GitHub repository at https://github.com/FARotiroti/Fourier, to demonstrate the Fourier approach's performance on a hierarchical model, we also provide an additional example, a multivariate g-and-k model, as depicted in Figure 3 of [9]. As in the univariate g-and-k example, in the multivariate extension, the Fourier approach provides estimates closely matched to the exact results. Because in Section 5.5, we will apply the Fourier approach to a model that involves both hierarchy and time-dependence, we will forgo further expansion on the hierarchical g-and-k model here and direct those interested to the aforementioned repository.

5.2. Univariate mixture

While the CvM performs reasonably well in the previous example, we show that even for a simple toy example in the ABC literature, the method falters. Introduced in [6], the standard toy example here considered involves a model given by

$$\theta \sim \mathcal{U}[-10, 10], \quad f(x|\theta) = 0.5\mathcal{N}(x|\theta, 1) + 0.5\mathcal{N}\left(x|\theta, 0.1^2\right).$$

We suppose that the given data is the single point $y = 0$. We run a Gibbs sampler using the Fourier likelihood in place of the true likelihood. The results appear in Figure 15. At each iteration of the sampler, the size of the replicated data sets is taken to be $N = 10^4$, for each of $M = 10$ parallel processes.

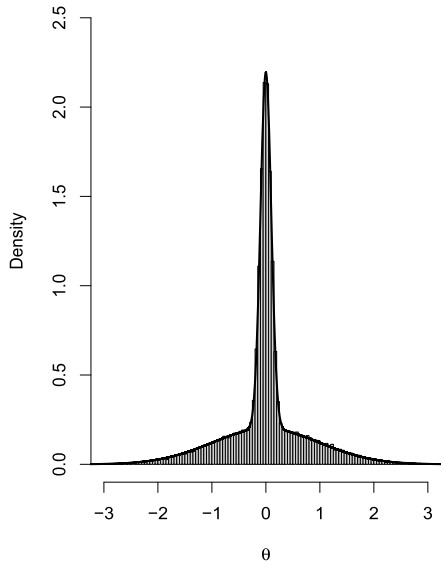

FIG 15. *Histogram of samples obtained from Fourier approach. The black line represents the target density, as computed on a grid.*

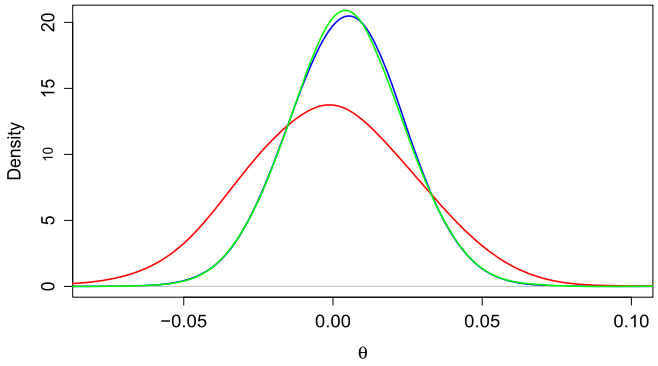

FIG 16. *Comparison of posterior densities of θ using the exact likelihood (blue), a Fourier likelihood (green), and CvM (red).*

Immediately, we recognize that CvM cannot be preformed with a single data point, since the distance evaluates identically with one observed data point. Conceding this point, let us then suppose that we take an i.i.d. sample of size $n = 100$ from the mixture. In Figure 16, we observe that while the Fourier approach is able to replicate the exact posterior closely, CvM simply concentrates around the given value of $\theta = 0$, revealing that it is not actually behaving properly as an estimator of the likelihood and so fails to represent the true posterior.

5.3. M/G/1 queue

Next, we consider the M/G/1 queuing model, a stochastic single-server queue model with Poisson arrivals, service times given by $\mathcal{U}(\theta_1, \theta_2)$, and service time between arrivals distributed as $\text{Exp}(\theta_3)$. The observed data are of length $n = 50$, consisting of the inter-departure times of 51 customers. The true parameters are set as $\theta = (1, 5, 0.2)$, and the prior is given by

$$\mathcal{U}\left(0, \min_i(y_i)\right) \times \mathcal{U}\left(0, 10 + \min_i(y_i)\right) \times \mathcal{U}\left(0, 0.5\right),$$

with $i = 1, \ldots, n$. For comparison, we again have access to the true posterior, here using the sampling scheme proposed by [26], which relies on data augmentation.

We repeat the experiment over 50 simulated data sets, using the same data as [10]. According to [10], the KDE (log) approach performs very well across all the comparisons; that is, KDE applied to log-transformed data.

For the three parameters, respectively, Tables 9–11 contain the repeated simulation results based on the 50 simulated data sets. For the Fourier approach, to set the parameters for the Yeo-Johnson transformation, we partition the

TABLE 9
Simulation results, based on 50 data sets, for parameter θ_1 of the M/G/1 queue example with simulated data of size $n = 50$.

Method	mean	median	std	80%	90%	95%
Exact	0.997	1.043	0.148	0.84	0.90	0.94
Fourier	0.978	1.022	0.154	0.82	0.88	0.94
Fourier (log)	0.983	1.029	0.152	0.84	0.88	0.94
KDE (log)	0.983	1.029	0.152	0.84	0.88	0.94

TABLE 10
Simulation results, based on 50 data sets, for parameter θ_2 of the M/G/1 queue example with simulated data of size $n = 50$.

Method	mean	median	std	80%	90%	95%
Exact	4.962	4.913	0.369	0.82	0.92	0.94
Fourier	5.010	4.976	0.519	0.80	0.90	0.98
Fourier (log)	5.071	5.001	0.544	0.78	0.90	0.96
KDE (log)	4.997	4.974	0.534	0.82	0.90	1.00

TABLE 11
Simulation results, based on 50 data sets, for parameter θ_3 of the M/G/1 queue example with simulated data of size $n = 50$.

Method	mean	median	std	80%	90%	95%
Exact	0.207	0.206	0.029	0.74	0.82	0.94
Fourier	0.198	0.198	0.022	0.74	0.82	0.90
Fourier (log)	0.202	0.201	0.025	0.74	0.82	0.86
KDE (log)	0.198	0.198	0.022	0.74	0.80	0.86

data based on empirical quartiles, as in Section 5.1. To compare more directly with KDE, we separately applied the Fourier approach to log-transformed data, with R then matched to the inverse bandwidth of KDE (log). We see that all the methods produce estimates closely matched to the exact values on average. Results for estimated posterior correlations can be found in the Appendix (Table 15).

In the next section, we will reveal the limitations of KDE, observing how the estimation suffers as the dimension of the problem increases. In addition, note the small observed data size of $n = 50$. As [10] observe, while this small size leads to difficulties in determining the underlying parameter values overall, its effect is most pronounced on the full-data ABC approaches; still, the Fourier approach, which also uses the full data, continues to provide highly accurate estimates.

5.4. Multivariate mixture

Extending the example in Section 5.2, we consider the multivariate model given by

$$\theta \sim \mathcal{U}(-10, 10)^3, \quad f(x|\theta) = 0.5\,\mathcal{N}(x|\theta, \Sigma) + 0.5\,N\left(x|\theta, 0.1^2\,\Sigma\right),$$

with observed data $y = 0$. We take

$$\Sigma = \begin{pmatrix} 13.98 & -1.64 & 0.44 \\ -1.64 & 2.63 & 1.74 \\ 0.44 & 1.74 & 4.07 \end{pmatrix}.$$

This matrix is a submatrix based on the five-dimensional covariance matrix in [23]. Note that this example is similar to the multivariate mixture considered in in [8], only there the covariance matrix is diagonal.

Figures 17–19 show simulation results based on using the exact likelihood, a Fourier likelihood, and KDE, with the latter two taking $N = 10^4$. First, we must point out that the computation using KDE was prohibitive for the original five-dimensional matrix, hence the reduction in dimension. Like CvM, KDE, in failing to approximate the likelihood properly, especially in higher dimensions, results in densities that deviate from the true posterior.

Thus, we recognize a salient benefit of the Fourier approach: There is no need to specify a covariance matrix for the likelihood estimation, as would be required in setting the bandwidth matrix for the multivariate normal kernel in KDE. The Fourier integral incorporates this information, and we are required only to set the R value for the given N, over which we also have control. Note that in Example 5.3, KDE performs very well, but it is worth keeping in mind its limitations. The full-data distance-based approaches also encounter problems in higher dimensions; as stated by [10], "Currently, [full data approaches] are predominantly suited to univariate data sets. We speculate that the application of these distances to univariate problems is due to both computational and statistical concerns."

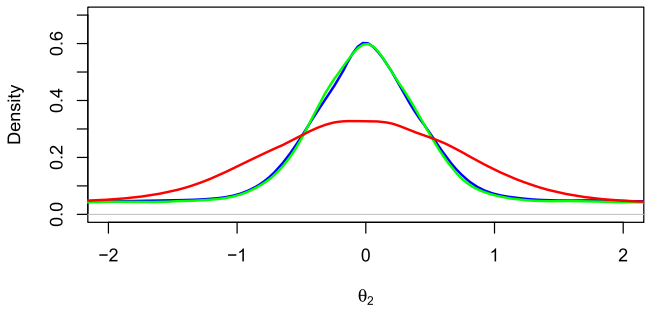

FIG 17. *Comparison of marginal posterior densities of θ_1 using the exact likelihood (blue), a Fourier likelihood (green), and KDE (red).*

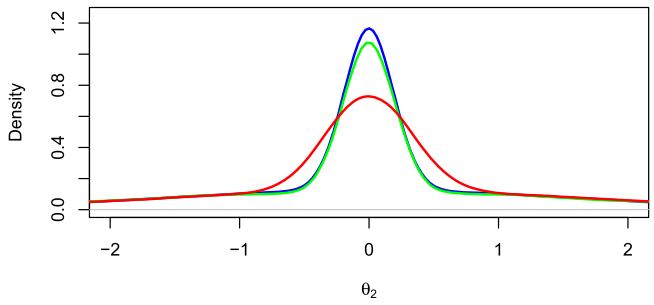

FIG 18. *Comparison of marginal posterior densities of θ_2 using the exact likelihood (blue), a Fourier likelihood (green), and KDE (red).*

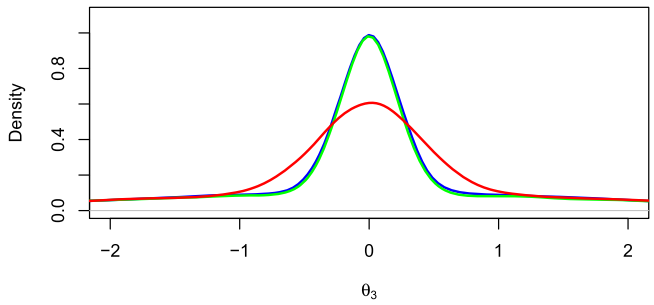

FIG 19. *Comparison of marginal posterior densities of θ_3 using the exact likelihood (blue), a Fourier likelihood (green), and KDE (red).*

The ability of the Fourier approach to handle multivariate data enables it to deal with time-dependent data, as the next example illustrates.

5.5. Ricker model

Before assessing the Ricker model, we address the fact that the data in the Ricker model are dependent. Owing to this dependence, many of the alternative approaches in the previous sections are outright impracticable or at least unreliable.

5.5.1. Dependent data

As with i.i.d. data, with correlated data, we can use knowledge of the structure of the data to our advantage. To illustrate, we consider data following an AR(1) process, with AR(1) parameter θ. That is, the model takes the form

$$y_{t+1} = \theta y_t + \epsilon_t,$$

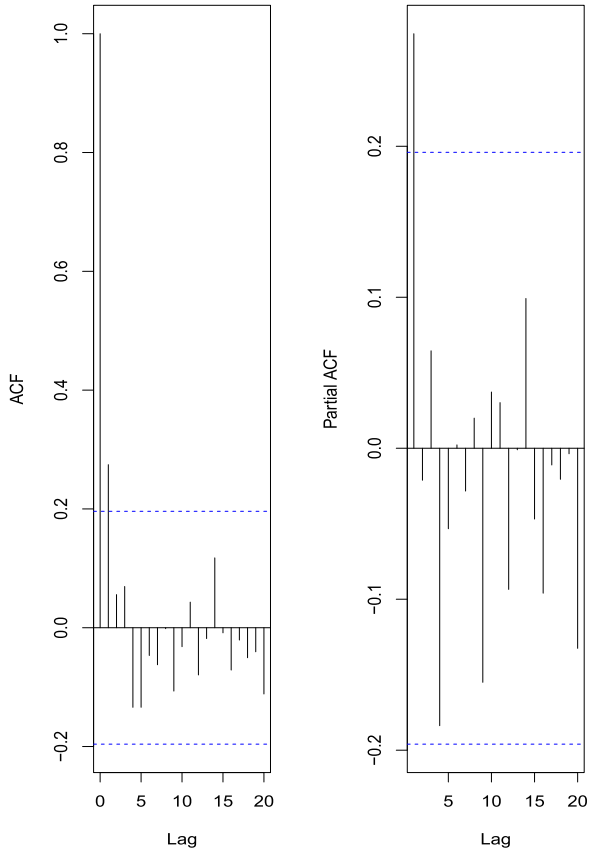

FIG 20. *Left: ACF of AR(1) data. Right: PACF of AR(1) data.*

where $\epsilon_t \sim_{iid} \mathcal{N}(0,1)$ and $y_0 \sim \mathcal{N}(0, 1/(1-\theta^2))$. From this model, we generate a series of length $n = 100$ with $\theta = 0.2$.

Acknowledging the dependence structure of the AR(1) process, as corroborated by the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the series, plotted in Figure 20, we can thus apply the Fourier approach to the successive dependent pairs. The densities of the posterior samples obtained from this scheme appear in Figure 21. At each iteration of the MCMC, the size of the replicated data sets is taken to be $N = 10^4$, for each of $M = 10$ parallel processes. The overlaid blue line represents the density of posterior samples obtained using the true AR(1) likelihood.

For a linear AR(p) process, $p \in \{1, 2, \dots\}$, of the form

$$y_t = \sum_{j=1}^{p} \theta_j \, y_{t-j} + \epsilon_t,$$

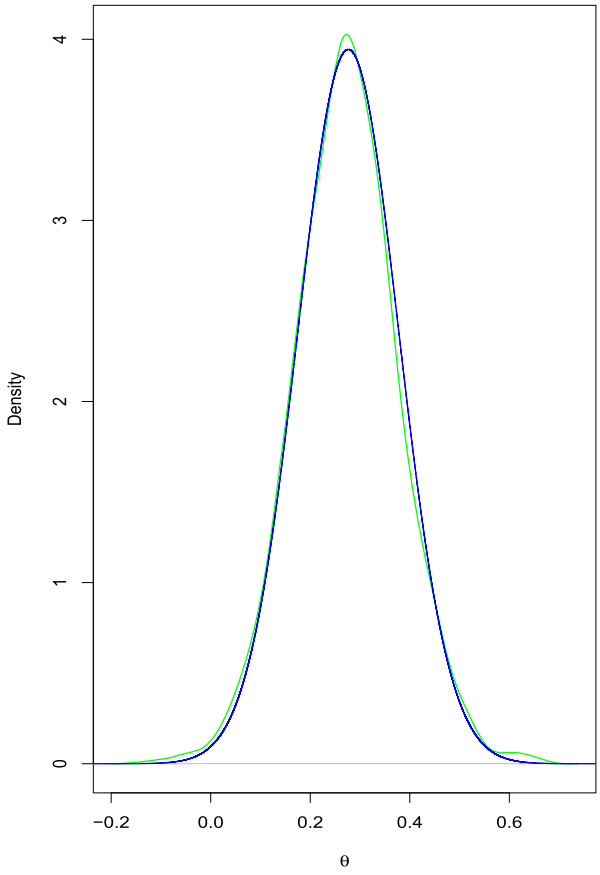

FIG 21. *The green line represents the density of posterior samples obtained from the Fourier approach, with* $N = 10^4$ *and* $M = 10$. *The blue line represents the density of samples obtained using the true likelihood function.*

where the (ϵ_t) are independent, one could also use the independent variables

$$z_t = y_t - \sum_{j=1}^{p} \theta_j \, y_{t-j}$$

when generating the pseudo-data. In the simplest case that $y_t = y_{t-1} + \epsilon_t$, the $z_t = y_t - y_{t-1}$, for $t = 1, 2, \ldots$, are independent.

Our approach to this problem is to estimate the likelihood using

$$\sum_{i=1}^{N} \prod_{l=1}^{p} k_R(y_l - x_{il}) \prod_{j=p+1}^{n} \frac{\sum_{i=1}^{N} \prod_{l=j-p}^{j} k_R(y_l - x_{il})}{\sum_{i=1}^{N} \prod_{l=j-p}^{j-1} k_R(y_l - x_{il})}, \tag{11}$$

where the (x_{il}) are simulated from the model for a particular parameter set and k_R is the Fourier kernel.

5.5.2. Ricker data

The hidden Markov model (HMM) is a fundamental model in which a latent state evolves according to Markovian dynamics, with corresponding noisy observations. Even an HMM with known dynamics can pose challenges for statistical inference [18]. For a non-linear HMM, for example,

$$y_t \sim f(\cdot \mid x_t), \quad x_t = g(x_{t-1}) + \eta_t,$$

to proceed with the Fourier approach, we would want to estimate the likelihood function

$$\prod_t p(y_t \mid y_{1:t-1}).$$

This estimation can be complicated with the Fourier approach owing to the large number of product terms which would need to appear in (3); however, if, as is usually the case, each $p(y_t \mid y_{1:t-1})$ can be adequately approximated by a p-order process, for some positive integer p, then we can use $p(y_t \mid y_{(t-p):(t-1)})$, and the Fourier approach now becomes highly attractive; from a mathematical perspective, $p(y_t \mid y_{(t-p):(t-1)})$ would be intractable, and yet, the Fourier approach estimates it with high accuracy.

Now that we have a better understanding of how to deal with dependent data, we can begin our analysis of data based on the scaled Ricker map, which has a hidden-Markov-model structure. Following [30], we consider the scaled Ricker map as prototypical of an ecological model with complex dynamics: It describes the dynamics of a population of size N at time t as follows:

$$N_{t+1} = rN_t \exp(-N_t + e_t), \quad [y_t \mid N_t] = \text{Pois}(\phi N_t),$$

where the $e_t \sim_{iid} \mathcal{N}(0, \sigma_e^2)$ denote process noise and r denotes the growth rate. The observed series (y_t) consists of Poisson random deviates with mean ϕN_t based on unobserved (N_t). Like [30], we are interested in making statistical inferences about $\theta = (\log r, \sigma_e, \phi)$, but as [30] observes, any likelihood-based inference about θ in this setting runs into issues of both analytical and numerical intractability.

Even though the proposed Fourier approach is not ideally suited to working directly with discrete data, this challenge can be easily overcome by simply adding a small amount of Gaussian noise to each value. Figure 22 shows the time series and histogram of the original Ricker data in the top and bottom left panels, respectively, and the time series and histogram of these data after a Yeo-Johnson transformation in the top and bottom right panels, respectively.

On the basis of Table 3.1 of [27], the ACF and PACF of the Ricker data, plotted in Figure 23, suggest that an eighth-order ($p = 8$) autoregressive model might provide a good fit to these data: the PACF cuts off distinctly after lag 8, while the ACF tails off.

The first row of Figure 24 shows the densities of the posterior samples for the three parameters of the Ricker model (namely, $\log r$, σ_e, and ϕ) obtained by applying the Fourier approach to the decomposed series, assuming an AR(8)

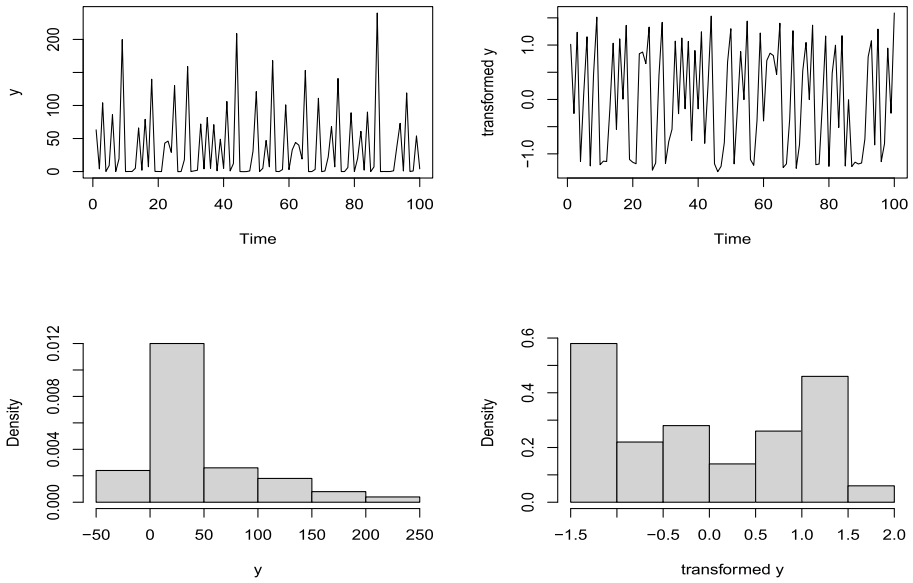

FIG 22. *Top-left: Time series of Ricker data. Top-right: Time series of transformed data. Bottom-left: Histogram of Ricker data. Bottom-right: Histogram of transformed data.*

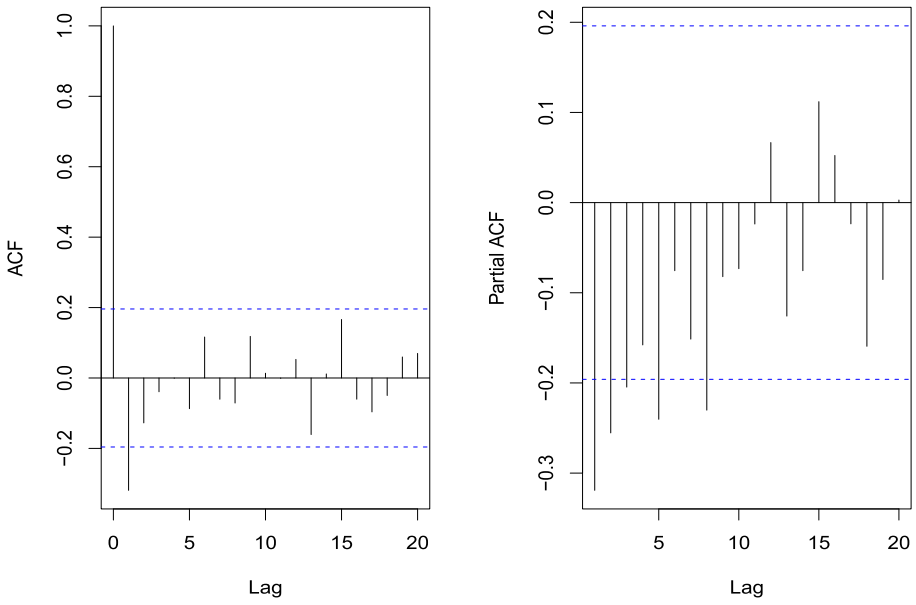

FIG 23. *Left: ACF of Ricker data. Right: PACF of Ricker data.*

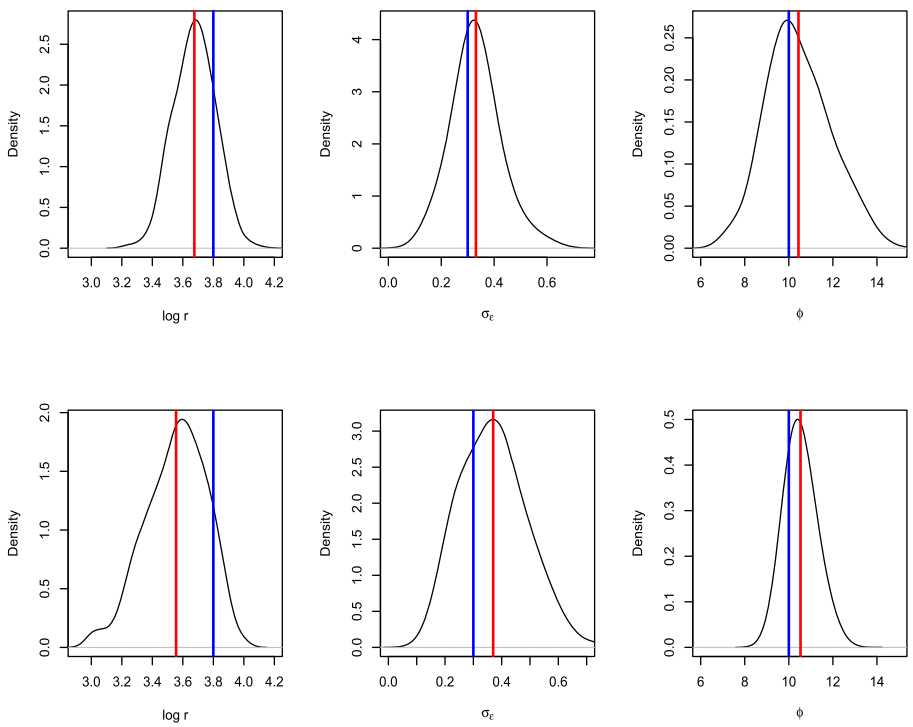

FIG 24. *Comparison of densities of posterior samples of the parameters of the Ricker model (namely, log r, σ_e, and ϕ). The first row is based on posterior samples obtained from the Fourier approach; the second row, from synthetic likelihood. The red lines indicate the mean of the samples, and the blue lines, the parameter value used to generate the data.*

structure and implementing (11). At each iteration of the MCMC, the size of the replicated data sets is taken to be $N = 10^4$, for each of $M = 10$ parallel processes. Since the dimension has increased, we take R in the range of 3 to 6 for the respective nine- and eight-dimensional numerators and denominators, composed of the Yeo-Johnson-transformed data. The motivation for these choices is to balance the bias and variance as the dimension increases. As before, the choice of R can be determined by following the procedure in [23], regardless of whether a normalizing transformation is used. Unlike SL, that is, the Fourier approach assumes a normalizing transformation for the purposes of setting the R value automatically with a scaled and centered Yeo-Johnson transformation, but it is equally valid to set the R value by observing the trace of the log-likelihood estimates across R values, a process facilitated by the linked code repository. From Figure 22, we note that although the transformation has helped to improve the symmetry of the data, they are still quite clearly non-normal. For SL, however, having normal summary statistics is a necessary part of the method, given its dependence on the multivariate normal log likelihood, as depicted in Figure 2.

TABLE 12
Posterior summaries for data simulated from Ricker model of size $n = 100$.

Par	True	Fourier			Synthetic likelihood		
		Mean	Median	SD	Mean	Median	SD
$\log r$	3.8	3.676	3.678	0.134	3.556	3.569	0.198
σ_e	0.3	0.319	0.331	0.088	0.370	0.366	0.113
ϕ	10	10.438	10.265	1.446	10.531	10.445	0.629

The second row of Figure 24 shows the result of applying SL. While both approaches agree on the general posterior structure and identify similar posterior mean values of the parameters, the associated mean and standard deviation of the samples recorded in Table 12 establish that overall the Fourier approach using the AR(8) decomposition identifies the true underlying parameter values with high accuracy. To validate these results, we ran a particle MCMC [16], which produced posterior means of 3.662, 0.317, and 10.298 for the respective parameters, output consistent with that of the Fourier approach.

Furthermore, confronted with any change to the model, SL would need to be carefully reassessed to determine the relevant statistics, whereas the Fourier approach could be applied with minimal changes. For instance, the Nicholson's blowfly model considered in [30] uses 16 summary statistics. With a given set of summary statistics, ABC can also potentially outperform SL when the statistics fail to be sufficiently normal.

Thus, the problem for SL and related ABC methods is essentially that of choosing summary statistics. Choosing summary statistics is difficult for models with complicated structures, like those with multiple levels of hierarchy, and requiring normality gives rise to further challenges. [11] outline a semi-automatic procedure for constructing appropriate summary statistics for ABC, using simulation to find estimates of the posterior means of the parameters, although Matteo Fasiolo, Natalya Pya and Simon N. Wood point out in the discussion of [11] that for the Ricker model, the chaotic nature of the dynamics might require a more sensitive treatment. Also note that the Fourier approach can be used with summary statistics and, in fact, can help to assess the distributional assumptions on these statistics, especially in higher dimensions.

We conclude by emphasizing that the Fourier approach exploits the same general scheme in all of the examples in this section, revealing its versatility and ease of implementation. Although a Yeo-Johnson transformation can be applied automatically, the Fourier approach can also accommodate whatever such transformation might be desired, and while transforming the data to appear more normal or at least symmetric can improve the estimation, Figure 22 reveals that the approach is robust to data that, even after transformation, appear quite clearly non-normal.

6. Conclusions

We have introduced a method for obtaining posterior samples from a model for which the likelihood function is intractable, or otherwise unavailable, but from

which simulation is possible. What distinguish this approach from alternatives, such as approximate Bayesian computation [17] and synthetic likelihood [30], are the ease of implementation, reliability of performance, and generality of application. Setting the two tuning parameters on which the Fourier approach relies, namely, R and N, is straightforward, particularly inasmuch as we have control over the sample size N of the replicated data, and the bias of the estimator can be of order $\exp(-R^2)$ for data that have a mixture of normal appearance, either as given or after transformation. With knowledge of the structure of the data, be that knowledge given or inferred, the Fourier approach obtains highly accurate results.

The Fourier approach further holds promise for Bayesian model choice. [22] advise caution with the use of ABC in model choice in the presence of insufficient summary statistics, which result in unknown loss of information. Their concern is for trustworthy approximations to the posterior probabilities of models, achievable only when all the data are involved in the approximation. As we have shown, across a variety of examples, the Fourier approach, which uses the full data, is able to approximate the posterior distribution accurately, suggesting that the Fourier approach could serve as a reliable tool in model selection.

Appendix: Proofs

Proof of Theorem 1

Here we find the mean and variance of the Fourier estimator for normal observations. To this end, let $R > 0$, and let $f(x)$ be a Gaussian, $f(x) \propto \exp(-b(x-c)^2)$, with fixed real number c and fixed positive real number b. Then, in relation to the proposed estimator, we define

$$I_1(R) = \int_{-\infty}^{\infty} \frac{\sin(R(y-x))}{y-x} \exp(-b(x-c)^2)dx,$$

for which the first derivative with respect to R is given by

$$\begin{aligned} I_1'(R) &= \int_{-\infty}^{\infty} \cos(R(y-x)) \exp(-b(x-c)^2)dx \\ &= \exp(iR(y-c)) \int_{-\infty}^{\infty} \frac{\exp(-iRu)}{2} \exp(-bu^2)du \\ &\quad + \exp(-iR(y-c)) \int_{-\infty}^{\infty} \frac{\exp(-iRu)}{2} \exp(-bu^2)du. \end{aligned}$$

Appealing to the Fourier transform, we see that

$$\begin{aligned} &\int_{-\infty}^{\infty} \exp(-iRx) \exp(-bx^2)dx \\ &\quad = 2\pi \int_{-\infty}^{\infty} \exp(-iR2\pi u) \exp(-b4\pi^2u^2)du \\ &\quad = 2\pi \mathcal{F}(\exp(-b4\pi^2u^2)) = \frac{2\pi \exp(-R^2/4b)}{\sqrt{4\pi b}} = \frac{\sqrt{\pi} \exp(-R^2/4b)}{\sqrt{b}}. \end{aligned}$$

This result is due to the Fourier transform of a Gaussian; i.e., the Fourier transform of a Gaussian, $\exp(-cx^2)$, is given by $\sqrt{\frac{\pi}{c}}\exp\left(\frac{-\pi^2R^2}{c}\right)$, and so, taking $c = 4b\pi^2$ yields the desired result.

Thus, we find that

$$\begin{aligned} I_1'(R) &= \left[\frac{\exp(iR(y-c)) + \exp(-iR(y-c))}{2}\right]\frac{\sqrt{\pi}\exp(-R^2/4b)}{\sqrt{b}} \\ &= \frac{\sqrt{\pi}\cos(R(y-c))\exp(-R^2/4b)}{\sqrt{b}}. \end{aligned}$$

Now, we can find the indefinite integral, relying again on the exponential formulation of the cosine function and recognizing the initial condition $I_1(0) = 0$. We conclude that

$$\begin{aligned} I_1(R) = \frac{\pi}{2}&\exp(-b(y-c)^2) \\ &\cdot\left[\operatorname{erf}\left(\frac{R-2ib(y-c)}{2\sqrt{b}}\right) + \operatorname{erf}\left(\frac{R+2ib(y-c)}{2\sqrt{b}}\right)\right]. \end{aligned}$$

For $X \sim \mathcal{N}(\mu, \sigma^2)$, we have that

$$\begin{aligned} I_1(R) = \frac{\pi}{2\sqrt{2\pi\sigma^2}}&\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \\ &\cdot\left[\operatorname{erf}\left(\frac{\sigma R - \frac{i(y-\mu)}{\sigma}}{\sqrt{2}}\right) + \operatorname{erf}\left(\frac{\sigma R + \frac{i(y-\mu)}{\sigma}}{\sqrt{2}}\right)\right], \end{aligned}$$

so that as $R \to \infty$, we have that

$$\lim_{R\to\infty} I_1(R) = \frac{\pi}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right).$$

And so, the bias of the estimator at arbitrary $y = y_0$ tends to 0 as $R \to \infty$.

Next, let us examine the variance of the estimator in this case. To begin, we define

$$I_2(R) = \int_{-\infty}^{\infty} \frac{\sin^2(R(y-x))}{(y-x)^2}\exp(-b(x-c)^2)dx,$$

for which the first derivative with respect to R is given by

$$I_2'(R) = \int_{-\infty}^{\infty} \frac{\sin(2R(y-x))}{y-x}\exp(-b(x-c)^2)dx,$$

and the second derivative by

$$\begin{aligned} I_2''(R) &= \int_{-\infty}^{\infty} 2\cos(2R(y-x))\exp(-b(x-c)^2)dx \\ &= 2\frac{\exp(i2R(y-c))}{2}\int_{-\infty}^{\infty}\exp(-i2Ru)\exp(-bu^2)du \\ &\quad + 2\frac{\exp(-i2R(y-c))}{2}\int_{-\infty}^{\infty}\exp(-i2Ru)\exp(-bu^2)du. \end{aligned}$$

Again appealing to the Fourier transform, we find that

$$I_2''(R) = 2\left[\frac{\exp(i2R(y-c)) + \exp(-i2R(y-c))}{2}\right] \cdot \frac{\sqrt{\pi}\exp(-R^2/b)}{\sqrt{b}}$$
$$= \frac{2\sqrt{\pi}\cos(2R(y-c))\exp(-R^2/b)}{\sqrt{b}}.$$

Taking the integral with respect to R, we find that

$$I_2'(R) = \frac{\pi}{2}\exp\big(-b(y-c)^2\big)\left[\operatorname{erf}\left(\frac{R-ib(y-c)}{\sqrt{b}}\right) + \operatorname{erf}\left(\frac{R+ib(y-c)}{\sqrt{b}}\right)\right].$$

And then, integrating again with respect to R, we find that

$$\begin{aligned} I_2(R) = {} & \frac{\pi}{2}\exp(-b(y-c)^2) \\ & \cdot \left\{\sqrt{\frac{b}{\pi}}\left[\exp\left(\frac{-[R+ib(y-c)]^2}{b}\right) + \exp\left(\frac{-[R-ib(y-c)]^2}{b}\right)\right]\right\} \\ & + \frac{\pi}{2}\exp(-b(y-c)^2)\left\{[R+ib(y-c)]\operatorname{erf}\left(\frac{R+ib(y-c)}{\sqrt{b}}\right)\right\} \\ & + \frac{\pi}{2}\exp(-b(y-c)^2)\left\{[R-ib(y-c)]\operatorname{erf}\left(\frac{R-ib(y-c)}{\sqrt{b}}\right)\right\} \\ & + C, \end{aligned}$$

with

$$C = -\sqrt{b\pi} - \operatorname{Re}\left(i\pi b(y-c)\exp(-b(y-c)^2)\operatorname{erf}(i\sqrt{b}(y-c))\right).$$

When $x = 0$ and $c = 0$, we find that

$$I_2(R) = \pi\left[R\cdot\operatorname{erf}\left(\frac{R}{\sqrt{b}}\right) + \sqrt{\frac{b}{\pi}}\exp\left(-\frac{R^2}{b}\right)\right] - \sqrt{b\pi}$$
$$= \pi R\cdot\operatorname{erf}\left(\frac{R}{\sqrt{b}}\right) + \sqrt{b\pi}\left[\exp\left(-\frac{R^2}{b}\right) - 1\right].$$

For $X \sim \mathcal{N}(\mu, \sigma^2)$, we have that

$$\begin{aligned} I_1(R,\mu,1/(2\sigma^2)) = {} & \frac{\pi}{2\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right) \\ & \cdot\left[\operatorname{erf}\left(\frac{\sigma R - \frac{i(y-\mu)}{\sigma}}{\sqrt{2}}\right) + \operatorname{erf}\left(\frac{\sigma R + \frac{i(y-\mu)}{\sigma}}{\sqrt{2}}\right)\right], \end{aligned} \tag{12}$$

and

$$
\begin{aligned}
&I_2(R,\mu,1/(2\sigma^2))\\
&\quad=\frac{\pi}{2\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)\\
&\quad\cdot\left\{\sqrt{\frac{1}{2\pi\sigma^2}}\left[\exp\left(-2\sigma^2\left[R-\frac{i}{2\sigma^2}(y-\mu)\right]^2\right)+\exp\left(-2\sigma^2\left[R+\frac{i}{2\sigma^2}(y-\mu)\right]^2\right)\right]\right\}\\
&\quad+\frac{\pi}{2\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)\left\{\left[R+\frac{i}{2\sigma^2}(y-\mu)\right]\operatorname{erf}\left(\frac{R+\frac{i}{2\sigma^2}(y-\mu)}{\sqrt{\frac{1}{2\sigma^2}}}\right)\right\}\\
&\quad+\frac{\pi}{2\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)\left\{\left[R-\frac{i}{2\sigma^2}(y-\mu)\right]\operatorname{erf}\left(\frac{R-\frac{i}{2\sigma^2}(y-\mu)}{\sqrt{\frac{1}{2\sigma^2}}}\right)\right\}\\
&\quad-\frac{1}{2\sigma^2}-\operatorname{Re}\left(i\frac{\sqrt{\pi}}{2\sigma^2\sqrt{2\sigma^2}}(y-\mu)\exp\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)\operatorname{erf}\left(i\sqrt{\frac{1}{2\sigma^2}}(y-\mu)\right)\right).
\end{aligned}
\tag{13}
$$

Proof of Theorem 2

With N fixed, consider

$$
\inf_{R>0}\{\mathrm{MSE}\left(\tilde{f}_{R,N}(y)\right)\}=\inf_{R>0}\{\mathrm{Var}(\tilde{f}_{R,N}(y))+[\mathrm{Bias}(\tilde{f}_{R,N}(y))]^2\}.
$$

In particular, for the standard normal $\mathcal{N}(0,1)$ at $x=0$, we have that

$$
I_1(R)|_{x=0}=\frac{\pi}{\sqrt{2\pi}}\operatorname{erf}\left(\frac{R}{\sqrt{2}}\right),
$$

and

$$
I_2(R)|_{x=0}=\frac{R\sqrt{\pi}}{\sqrt{2}}\operatorname{erf}\left(R\sqrt{2}\right)+\frac{1}{2}\left[\exp\left(-2R^2\right)-1\right].
$$

Thus, setting $f(y)=\mathcal{N}(0,1)$ and $y=0$, we find that

$$
\begin{aligned}
\mathrm{MSE}\left(\tilde{f}_{R,N}(0)\right)&=\frac{1}{N\pi^2}\left(I_2(R)-[I_1(R)]^2\right)+\left[\frac{1}{\pi}I_1(R)-\mathcal{N}(0|0,1)\right]^2\\
&=\frac{1}{N\pi^2}\left(\frac{R\sqrt{\pi}}{\sqrt{2}}\operatorname{erf}\left(R\sqrt{2}\right)+\frac{1}{2}\left[\exp\left(-2R^2\right)-1\right]-\frac{\pi}{2}\operatorname{erf}^2\left(\frac{R}{\sqrt{2}}\right)\right)\\
&\quad+\left[\frac{1}{\sqrt{2\pi}}\operatorname{erf}\left(\frac{R}{\sqrt{2}}\right)-\frac{1}{\sqrt{2\pi}}\right]^2.
\end{aligned}
\tag{14}
$$

To find the desired infimum, we differentiate the MSE with respect to R:

$$\begin{aligned}
&\frac{d}{dR}\mathrm{MSE}\left(\tilde{f}_{R,N}(0)\right)\\
&\quad=\frac{1}{N\pi^2}\left(\frac{R}{\sqrt{2\pi}}\mathrm{erf}\left(R\sqrt{2}\right)+\frac{1}{2\pi}\left[\exp\left(-2R^2\right)-1\right]-\frac{1}{2\pi}\mathrm{erf}^2\left(\frac{R}{\sqrt{2}}\right)\right)\\
&\qquad+\left[\frac{1}{\sqrt{2\pi}}\mathrm{erf}\left(\frac{R}{\sqrt{2}}\right)-\frac{1}{\sqrt{2\pi}}\right]^2\\
&\quad=\frac{\exp(-R^2/2)\left[2(N-1)\mathrm{erf}(R/\sqrt{2})+\exp(R^2/2)\mathrm{erf}(R\sqrt{2})-2N\right]}{N\sqrt{2}\pi^{3/2}}.
\end{aligned}$$

Proof of Corollary 1

As $N\to\infty$, we have that $\inf_{R>0}\mathrm{MSE}\left(\tilde{f}_{R,N}(0)\right)$ occurs where

$$2\left[\frac{1}{\sqrt{2\pi}}\mathrm{erf}\left(\frac{R}{\sqrt{2}}\right)-\frac{1}{\sqrt{2\pi}}\right]\left[\frac{1}{\sqrt{2\pi}}\frac{2}{\sqrt{\pi}}\exp\left(-\frac{R^2}{2}\right)\right]\equiv 0;$$

i.e., $\mathrm{erf}\left(R/\sqrt{2}\right)=1$, implying that, given the machine precision of R (the programming language) and the numerical approximation to the erf function due to [1], the approximation error is vanishingly small for $R\geq 5.923\cdot\sqrt{2}$. Accordingly, we optimally set $R\approx 5.923\cdot\sqrt{2}$.

Proof of Corollary 2

Unbiasedness is due to $\max(0,a)=(a+|a|)/2$, in conjunction with Definition 4 and the i.i.d. $\{x_{ik}\}$.

Additional results

The results in Tables 13, 14 and 15 are for estimated posterior correlations for the g-and-k distribution and M/G/1 queue examples.

TABLE 13
Average correlation, based on 50 data sets, between parameters of the g-and-k distribution example with simulated data of size $n=100$.

Method	$\mathrm{corr}(a,b)$	$\mathrm{corr}(a,g)$	$\mathrm{corr}(a,k)$	$\mathrm{corr}(b,k)$	$\mathrm{corr}(b,g)$	$\mathrm{corr}(g,k)$
Exact	0.691	−0.397	−0.532	0.136	−0.724	0.123
Fourier	0.687	−0.413	−0.528	0.112	−0.722	0.136
CvM	0.644	−0.454	−0.446	0.070	−0.631	0.327

TABLE 14
Average correlation, based on 50 data sets, between parameters of the g-and-k distribution example with simulated data of size $n = 1000$.

Method	$\text{corr}(a, b)$	$\text{corr}(a, g)$	$\text{corr}(a, k)$	$\text{corr}(b, k)$	$\text{corr}(b, g)$	$\text{corr}(g, k)$
Exact	0.748	−0.329	−0.610	0.166	−0.764	0.173
Fourier	0.751	−0.327	−0.611	0.165	−0.762	0.178
CvM	0.720	−0.392	−0.536	0.078	−0.669	0.413

TABLE 15
Average correlation, based on 50 data sets, between parameters of the M/G/1 queue example with simulated data of size $n = 50$.

Method	$\text{corr}(\theta_1, \theta_2)$	$\text{corr}(\theta_1, \theta_3)$	$\text{corr}(\theta_2, \theta_3)$
Exact	−0.020	−0.004	0.013
Fourier	−0.018	−0.068	−0.035
Fourier (log)	−0.018	−0.061	−0.019
KDE (log)	−0.021	−0.062	−0.046

Acknowledgments

The authors are grateful for the comments and suggestions of two referees on an earlier version of the paper.

References

[1] ABRAMOWITZ, M. and STEGUN, I. A. (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover. MR0415962

[2] ALQUIER, P., FRIEL, N., EVERITT, R. and BOLAND, A. (2016). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing* **26** 29–47. MR3439357

[3] ANDRIEU, C., LEE, A. and VIHOLA, M. (2018). Theoretical and methodological aspects of MCMC computations with noisy likelihoods. *Chapman & Hall/CRC Handbooks of Modern Statistical Methods*. MR3889286

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

[5] BEAUMONT, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics* **164** 1139–1160.

[6] BEAUMONT, M. A., CORNUET, J.-M., MARIN, J.-M. and ROBERT, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika* **96** 983–990. MR2767283

[7] BERNTON, E., JACOB, P. E., GERBER, M. and ROBERT, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *Jour-*

nal of the Royal Statistical Society Series B: Statistical Methodology **81** 235–269. MR3928142

[8] Bonassi, F. V. and West, M. (2015). Sequential Monte Carlo with adaptive weights for approximate Bayesian computation. *Bayesian Analysis* **10** 171–187. arXiv:1503.07791. https://doi.org/10.1214/14-BA891. MR3420901

[9] Clarté, G., Robert, C. P., Ryder, R. J. and Stoehr, J. (2021). Componentwise approximate Bayesian computation via Gibbs-like steps. *Biometrika* **108** 591–607. MR4298766

[10] Drovandi, C. and Frazier, D. T. (2022). A comparison of likelihood-free methods with and without summary statistics. *Statistics and Computing* **32** 1–23. MR4426803

[11] Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 419–474. MR2925370

[12] Frazier, D. T. (2020). Robust and efficient approximate Bayesian computation: A minimum distance approach. *arXiv preprint arXiv:2006.14126*.

[13] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109. MR3363437

[14] Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10** 25–37.

[15] Jiang, B. (2018). Approximate Bayesian computation with Kullback-Leibler divergence as data discrepancy. In *International conference on artificial intelligence and statistics* 1711–1721. PMLR.

[16] King, A. A., Nguyen, D. and Ionides, E. L. (2016). Statistical Inference for Partially Observed Markov Processes via the R Package pomp. *Journal of Statistical Software* **69**.

[17] Marin, J.-M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing* **22** 1167–1180. MR2992292

[18] Mattila, R., Rojas, C., Krishnamurthy, V. and Wahlberg, B. (2017). Inverse filtering for hidden Markov models. *Advances in Neural Information Processing Systems* **30**.

[19] Mitrovic, J., Sejdinovic, D. and Teh, Y. W. (2016). DR-ABC: Approximate Bayesian computation with kernel-based distribution regression. *Proceedings of Machine Learning Research* **48** 1482-1491.

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

[21] Park, M., Jitkrittum, W. and Sejdinovic, D. (2016). K2-ABC: Approximate Bayesian computation with kernel embeddings. In *Artificial intelligence and statistics* 398–407. PMLR.

[22] Robert, C. P., Cornuet, J.-M., Marin, J.-M. and Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences* **108** 15112–15117.

[23] Rotiroti, F. and Walker, S. G. (2022). Computing marginal likelihoods via the Fourier integral theorem and pointwise estimation of posterior densities. *Statistics and Computing* **32** 67. MR4470289

[24] Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics* 1151–1172. MR0760681

[25] Rubio, F. J. and Johansen, A. M. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics* **7** 1632-1654. MR3070873

[26] Shestopaloff, A. Y. and Neal, R. M. (2014). On Bayesian inference for the M/G/1 queue with efficient MCMC sampling. *arXiv preprint arXiv: 1401.5548*.

[27] Shumway, R. H. and Stoffer, D. S. (2017). *Time Series Analysis and Its Applications: With R Examples*, 4th ed. Springer. MR3642322

[28] Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge. MR0848134

[29] Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81** 82–86. MR0830567

[30] Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature* **466** 1102–1104.

[31] Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* **87** 954–959. MR1813988