An extended empirical saddlepoint approximation for intractable likelihoods

Matteo Fasiolo Simon N. Wood

School of Mathematics, University of Bristol,
Bristol BS8 1TH, United Kingdom
e-mail: matteo.fasiolo@bristol.ac.uk*;* simon.wood@bristol.ac.uk

Florian Hartig

Faculty of Biology and Pre-Clinical Medicine,
University of Regensburg, Regensburg 93053, Germany
e-mail: florian.hartig@biologie.uni-regensburg.de

and

Mark V. Bravington

CSIRO Marine Lab, 7004 Hobart, Australia
e-mail: Mark.Bravington@csiro.au

Abstract: The challenges posed by complex stochastic models used in computational ecology, biology and genetics have stimulated the development of approximate approaches to statistical inference. Here we focus on Synthetic Likelihood (SL), a procedure that reduces the observed and simulated data to a set of summary statistics, and quantifies the discrepancy between them through a synthetic likelihood function. SL requires little tuning, but it relies on the approximate normality of the summary statistics. We relax this assumption by proposing a novel, more flexible, density estimator: the Extended Empirical Saddlepoint approximation. In addition to proving the consistency of SL, under either the new or the Gaussian density estimator, we illustrate the method using three examples. One of these is a complex individual-based forest model for which SL offers one of the few practical possibilities for statistical inference. The examples show that the new density estimator is able to capture large departures from normality, while being scalable to high dimensions, and this in turn leads to more accurate parameter estimates, relative to the Gaussian alternative. The new density estimator is implemented by the `esaddle` R package, which is freely available on the Comprehensive R Archive Network (CRAN).

Keywords and phrases: Intractable likelihood, saddlepoint approximation, synthetic likelihood, simulation-based inference, implicit statistical model, density estimation.

Contents

1. Introduction

Synthetic Likelihood (SL) (Wood, 2010) is a simulation-based inferential proce-
dure similar to Approximate Bayesian Computation (ABC) (Beaumont et al.,
2002), but with the practical advantage of requiring less tuning. In Wood (2010)
tuning is avoided partly through a Gaussian assumption on the distribution of
the statistics used to compare the data and the model output. This assumption
can be problematic, as illustrated by the following simple population dynamic

model of an organism subject to boom and bust dynamics with stochastic external recruitment

$$N_{t+1} \sim \begin{cases} \text{Pois}\{N_t(1+r)\} + \epsilon_t, & \text{if } N_t \leq \kappa, \\ \text{Binom}(N_t, \alpha) + \epsilon_t, & \text{if } N_t > \kappa, \end{cases} \tag{1}$$

where $\epsilon_t \sim \text{Pois}(\beta)$ is a stochastic arrival process, with rate $\beta > 0$, and $t = 1, \ldots, \text{T}$. The population N_t grows stochastically at rate $r > 0$, but it crashes if the carrying capacity κ is exceeded. The severity of the crash depends on the survival probability $\alpha \in (0, 1)$. Two fairly natural statistics for this model are the population mean and the number of periods during which $N_t \leq 1$, the latter being useful for identifying β. However, as Figure 1 shows, the distribution of these statistics is far from normal, which could affect the accuracy of the parameter estimates produced by SL. The purpose of this paper is to develop a version of SL that relaxes the normality requirement, while retaining the tuning free advantages of the original method. We do this by replacing the Gaussian assumption with a new density estimator: the Extended Empirical Saddlepoint (EES) estimator. We prove the consistency of the resulting parameter estimator, and illustrate that the method can yield substantial inferential improvements when multivariate gaussianity is untenable. We also provide examples where ABC methods would require exceedingly low tolerances and low acceptance rates in order to achieve equivalent accuracy.

The most important commonality between SL and ABC methods is that both base statistical inference on a vector of summary statistics, $\boldsymbol{s}^0 = S(\boldsymbol{y}^0)$, rather than on the full data, $\boldsymbol{y}^0$. However, while ABC methods explicitly aim at sampling from an approximation to $p(\boldsymbol{\theta}|\boldsymbol{s}^0)$, SL provides a parametric approximation to $p(\boldsymbol{s}^0|\boldsymbol{\theta})$ (Hartig et al., 2011). This synthetic likelihood, $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$, can then be used within a Bayesian or a classical context. Wood (2010) used a multivariate Gaussian density to approximate the distribution of the summary statistics. Under this distributional assumption, a pointwise estimate of the synthetic likelihood at $\boldsymbol{\theta}$ can be obtained using Algorithm 1.

Algorithm 1 Estimating $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$

1: Simulate datasets $\boldsymbol{Y}_i, \ldots, \boldsymbol{Y}_m$ from the model $p(\boldsymbol{y}|\boldsymbol{\theta})$.
2: Transform each dataset $\boldsymbol{Y}_i$ to a vector of summary statistics $\boldsymbol{S}_i = S(\boldsymbol{Y}_i)$.
3: Calculate sample mean $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ of the simulated statistics, possibly robustly.

4: Estimate the synthetic likelihood

$$\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}) = (2\pi)^{-\frac{d}{2}} \det(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}})^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}})^T \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}) \right\},$$

where d is the number of summary statistics used.

The approximate posterior sampled by ABC algorithms is often of the form

$$p_\epsilon(\boldsymbol{\theta}|\boldsymbol{s}^0) \propto p_\epsilon(\boldsymbol{s}^0|\boldsymbol{\theta})p(\boldsymbol{\theta}) = \left[\int \mathbb{I}\{\rho(\boldsymbol{s}^0, \boldsymbol{s}) < \epsilon\} p(\boldsymbol{s}|\boldsymbol{\theta}) d\boldsymbol{s} \right] p(\boldsymbol{\theta}),$$

where $\epsilon > 0$ is the so-called tolerance or acceptance threshold, $\rho()$ is a distance function and $\mathbb{I}()$ is an indicator function. When $p_\epsilon(\boldsymbol{\theta}|\boldsymbol{s}^0)$ is sampled using Markov chain Monte Carlo (MCMC) methods, ABC can be viewed as a pseudo-marginal method (Andrieu et al., 2009) where the ABC likelihood is substituted by the unbiased estimator

$$\hat{p}_\epsilon(\boldsymbol{s}^0|\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\{\rho(\boldsymbol{s}^0, \boldsymbol{S}_i|\boldsymbol{\theta}) < \epsilon\},$$

with $\boldsymbol{S}_1,\ldots,\boldsymbol{S}_m \sim p(\boldsymbol{s}|\boldsymbol{\theta})$. The choice of tolerance is critical to the performance of ABC, as decreasing ϵ makes the ABC posterior closer to $p(\boldsymbol{\theta}|\boldsymbol{s}^0)$, but decreases the probability that any of the statistics simulated from the model will fall within distance ϵ of $\boldsymbol{s}^0$. A common choice of distance function is $\rho(\boldsymbol{s}^0, \boldsymbol{s}) = (\boldsymbol{s}^0 - \boldsymbol{s})^T \boldsymbol{Q}(\boldsymbol{s}^0 - \boldsymbol{s})$, where $\boldsymbol{Q}$ is a positive definite scaling matrix, needed to take into account the fact that the elements of $\boldsymbol{s}$ might be on different scales and possibly highly correlated. Given that the distribution of the summary statistics depends on $\boldsymbol{\theta}$, a single scaling matrix might not be sufficient for normalizing summary statistics simulated using different parameter values. See Fasiolo and Wood (2015) for an example where using a scaling matrix which does not depend on $\boldsymbol{\theta}$ leads to inaccurate ABC estimates, and Prangle et al. (2017) for an adaptive solution to this problem.

One advantage of SL, over most ABC methods, is that it does not require the user to choose ϵ and that the summary statistics are scaled automatically and dynamically by $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$. In addition, Blum (2010) showed that the convergence rate of ABC methods degrades rapidly with d. This curse of dimensionality, brought about by the non-parametric nature of ABC, forces practitioners to use dimension reduction or statistics selection techniques, such as those described by Blum et al. (2013). SL is less sensitive to the number of statistics used, due to the parametric likelihood approximation (Price et al., 2017).

The development of approximate inferential approaches, such as SL and ABC, has been driven by the increasing availability of computational resources and by the challenges to model-based inference emerging in computational biology, ecology and genetics. These methods address the issue that, for many scientifically motivated models, the likelihood function is intractable; it may be too expensive to evaluate, unknown, or too time-consuming to derive analytically. Furthermore, even when sophisticated integration approaches, such as particle filters (Doucet et al., 2000), could provide consistent likelihood estimators, using approximate methods might still be preferable in practice, because automation and robustness to implementation details. Particle filters often rely on the specific structure of the chosen model, hence their implementation may need substantial changes if a different model is considered. In contrast, SL and ABC treat the model as a black box, thus allowing practitioners to rapidly explore a variety of models. Approximate methods might also have computational advantages over particle filters, but this depends on factors such as the specific model and summary statistics considered. For example, the first two steps of Algorithm 1 can be computed in an embarrassingly parallel fashion, while parallelizing the
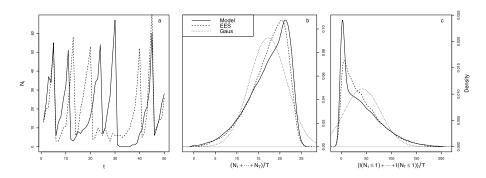
FIG 1. *a: First 50 steps of two sample paths, simulated from model (1). b and c: Marginal distributions of the population mean and of the number of periods t during which $N_t \leq 1$, when $T = 250$. Both distributions are highly skewed, and the EES density achieves a much better fit than a Gaussian approximation.*

resampling step of particle filters is less straightforward (see e.g. Murray et al. (2016)). However, the time needed to compute the summary statistics might more than offset this advantage.

The performance of SL, ABC and particle filters has been compared in detail by Fasiolo et al. (2016) and Everitt et al. (2015), respectively in the context of parameter estimation for non-linear state space models and of Bayesian model selection for Markov random field models. It is not the purpose of this paper to provide another extensive comparison. Instead, we focus on SL, and we start from the observation that the above-mentioned properties of this method are not without cost. In fact, although the Central Limit Theorem (CLT) assures asymptotic normality of many classes of statistics, improving the quality of the multivariate normal approximation is not easy in practice. Finding a suitable normalizing transformation is particularly challenging in the context of parameter estimation, because such transformation would need to ensure normality across the parameter space. This motivates the main contribution of this work: we relax the multivariate normality assumption, while maintaining the ease-of-use and scalability of SL. We achieve this by proposing a flexible density estimator, namely the EES approximation. In addition to illustrating empirically that, when the distribution of the summary statistics is far from normal, the EES-based version of SL leads to more accurate parameter estimates than its Gaussian alternative, we prove that maximizing the synthetic likelihood produces consistent parameter estimators, under either the EES or the Gaussian density estimator.

The paper is organized as follows. We introduce the empirical saddlepoint approximation in Section 2 and we propose its extended version in Section 3. In Section 4 we clarify how the new density estimator can be used within the context of SL and we prove the consistency of the resulting parameter estimator. In Sections 5 and 6 we illustrate the method on model (1) and on another

simple example, designed to show the potential limitations of ABC and of the Gaussian version of SL, and in Section 7 we apply the method to inference for a complex individual-based forest model, for which statistical inference is challenging without the use of summary statistics, while the model is sufficiently computationally costly that extensive method tuning is impractical.

2. Saddlepoint approximations

Recall that we are interested in using saddlepoint methods to closely approximate the statistics-based likelihood, $p(\boldsymbol{s}^0|\boldsymbol{\theta})$. However, the following discussion is valid beyond the context of SL, hence we temporarily suppress all dependencies on $\boldsymbol{\theta}$. We restore them in Section 4, which describes how the proposed density estimator can be used within SL.

We were led to saddlepoint approximations, among other multivariate density estimators, by the following considerations. While saddlepoint approximations are derived from asymptotic expansions, they are often very accurate even in small samples and, in contrast to Edgeworth approximations, they are strictly positive and do not show polynomial-like waves in the tails. In addition, their empirical version provides a close approximation to the density of widely used statistics, such as M- (Ronchetti and Welsh, 1994) and L-estimators (Easton and Ronchetti, 1986).

Saddlepoint expansions were introduced into the statistical literature by Daniels (1954) and can be used to approximate the density function of a random variable, starting from its moment or cumulant generating function. When $\boldsymbol{S}$ is a continuous d-dimensional random vector, its probability density function, $p(\boldsymbol{s})$, is associated with the moment generating function

$$M(\boldsymbol{\lambda}) = \mathbb{E}\big(e^{\boldsymbol{\lambda}^T \boldsymbol{S}}\big) = \int_{-\infty}^{+\infty} e^{\boldsymbol{\lambda}^T \boldsymbol{s}} p(\boldsymbol{s}) \, d\boldsymbol{s},$$

while the cumulant generating function is defined as $K(\boldsymbol{\lambda}) = \log M(\boldsymbol{\lambda})$. In the following we assume that $M(\boldsymbol{\lambda})$ exists for $\boldsymbol{\lambda} \in I$, where I is a non-vanishing subset of $\mathbb{R}^d$ containing the origin. The Taylor series coefficient of $M(\boldsymbol{\lambda})$ and $K(\boldsymbol{\lambda})$, that is the collection of higher derivatives evaluated at $\boldsymbol{\lambda} = \boldsymbol{0}$, are respectively the moments and cumulants of the random variable $\boldsymbol{S}$. It is easy to show that the first two cumulants are equal to the mean vector and covariance matrix of $\boldsymbol{S}$, that is $K'(\boldsymbol{0}) = \mathbb{E}(\boldsymbol{S}) = \boldsymbol{\mu}$ and $K''(\boldsymbol{0}) = \mathbb{E}\{(\boldsymbol{S} - \boldsymbol{\mu})(\boldsymbol{S} - \boldsymbol{\mu})^T\} = \boldsymbol{\Sigma}$. If $\boldsymbol{S}$ is a discrete random vector, the generating functions are obtained by substituting the integrals with summations over the support of $\boldsymbol{S}$.

Saddlepoint approximations rely on the one-to-one correspondence between the cumulant generating function and the probability density function of $\boldsymbol{S}$. For a continuous $\boldsymbol{S}$, the saddlepoint density is

$$\hat{p}(\boldsymbol{s}) = \frac{1}{(2\pi)^{\frac{d}{2}} \det\{K''(\hat{\boldsymbol{\lambda}})\}^{\frac{1}{2}}} e^{K(\hat{\boldsymbol{\lambda}}) - \hat{\boldsymbol{\lambda}}^T \boldsymbol{s}} , \tag{2}$$

where $\hat{\boldsymbol{\lambda}}$ is such that

$$K'(\hat{\boldsymbol{\lambda}}) = \boldsymbol{s}. \tag{3}$$

Condition (3) is often called the saddlepoint equation. When the solution $\hat{\boldsymbol{\lambda}}$ can be found analytically, then it is plugged into (2) which leads to an analytic saddlepoint density approximation. If the solution is not available analytically, then the saddlepoint approximation is implicit and must evaluated by solving (3) numerically (typically using Newton-Raphson algorithm) and then plugging the resulting numerical solution $\hat{\boldsymbol{\lambda}}$ into (2).

The saddlepoint density is defined only on the interior, $J_{V_{\boldsymbol{s}}}$, of the support, $V_{\boldsymbol{s}}$, of the original density, $p(\boldsymbol{s})$. Another important property of $\hat{p}(\boldsymbol{s})$ is that it is generally improper. A proper density can be obtained through normalization

$$\bar{p}(\boldsymbol{s}) = \frac{\hat{p}(\boldsymbol{s})}{\int_{J_{V_{\boldsymbol{s}}}} \hat{p}(\boldsymbol{s})d\boldsymbol{s}}.$$

While Durbin (1980) provides asymptotic arguments demonstrating the superior accuracy achieved by normalized saddlepoint approximations, the normalizing constant is often analytically unavailable and must be approximated numerically, which can be expensive when d is moderately large and the saddlepoint equation must be solved numerically. For a discrete $\boldsymbol{S}$ the saddlepoint expression (2) still applies and can be computed across $J_{V_{\boldsymbol{s}}}$, however its value can be interpreted as an approximation to $\mathrm{pr}(\boldsymbol{S} = \boldsymbol{s})$ only when all the elements of $\boldsymbol{S}$ are integer-valued. Formula (2) also applies to mixed settings, where some of the components of $\boldsymbol{S}$ are continuous, while others are discrete.

Saddlepoint approximations are typically more accurate than Gaussian approximations based on the CLT. Intuitively, this is because CLT approximations use only the first two cumulants of the distribution, while saddlepoint methods take into account all cumulants, hence the full shape of the cumulant generating function over its convergence region. This means that saddlepoint approximations use more information regarding the distribution, which often leads to higher accuracy. For a comprehensive introduction to saddlepoint approximations, see Butler (2007). In particular, Section 5.2 therein provides a derivation of (2) that does not requires any knowledge of complex analysis and that clarifies how saddlepoint methods relate to Edgeworth and Gaussian approximations.

2.1. Empirical saddlepoint approximation

Suppose that the analytic form of $K(\boldsymbol{\lambda})$ is unknown, as it generally is for simulation-based methods such as SL. If we can simulate from $p(\boldsymbol{s})$, then it is possible to estimate $K(\boldsymbol{\lambda})$ using the estimator proposed by Davison and Hinkley (1988)

$$\hat{K}_m(\boldsymbol{\lambda}) = \log \hat{M}_m(\boldsymbol{\lambda}) = \log\left(\frac{1}{m}\sum_{i=1}^{m} e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i}\right), \tag{4}$$

where m is the number of simulations used. Derivative estimates of $\hat{K}_m(\boldsymbol{\lambda})$ are

$$\hat{K}'_m(\boldsymbol{\lambda}) = \frac{\sum_{i=1}^{m} e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i} \boldsymbol{s}_i}{\sum_{i=1}^{m} e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i}}, \qquad \hat{K}''_m(\boldsymbol{\lambda}) = \frac{\sum_{i=1}^{m} e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i} \boldsymbol{s}_i \boldsymbol{s}_i^T}{\sum_{i=1}^{m} e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i}} - \hat{K}'_m(\boldsymbol{\lambda}) \hat{K}'_m(\boldsymbol{\lambda})^T.$$

These can be used to obtain an empirical saddlepoint approximation

$$\hat{p}_m(\boldsymbol{s}) = \frac{1}{(2\pi)^{\frac{d}{2}} \det\{\hat{K}''_m(\hat{\boldsymbol{\lambda}}_m)\}^{\frac{1}{2}}} e^{\hat{K}_m(\hat{\boldsymbol{\lambda}}_m) - \hat{\boldsymbol{\lambda}}_m^T \boldsymbol{s}}, \tag{5}$$

where $\hat{\boldsymbol{\lambda}}_m$ is the solution of

$$\hat{K}'_m(\hat{\boldsymbol{\lambda}}_m) = \boldsymbol{s}. \tag{6}$$

Notice that $\hat{K}'_m(\boldsymbol{\lambda})$ is a convex combination of the simulated vectors $\boldsymbol{s}_i$, hence (6) has no solution if $\boldsymbol{s}$ falls outside the convex hull of the $\boldsymbol{s}_i$'s. This limitation is addressed in Section 3.

Feuerverger (1989) provides asymptotic results regarding how well $\hat{p}_m(\boldsymbol{s})$ approximates $\hat{p}(\boldsymbol{s})$ in a univariate setting. In Appendix B.1 we show how these carry over to the current multivariate setting. In particular, $\hat{p}_m(\boldsymbol{s})$ converges to $\hat{p}(\boldsymbol{s})$ at parametric rate $O(m^{-1/2})$ for $\boldsymbol{\lambda} \in I/2$, where $I/2$ is the subset of I such that $\boldsymbol{\lambda} \in I/2$ if $2\boldsymbol{\lambda} \in I$, while the convergence is slower outside this region. Regardless of the distribution of $\boldsymbol{S}$, $\boldsymbol{s} = \boldsymbol{\mu} = \mathbb{E}(\boldsymbol{S})$ corresponds to $\boldsymbol{\lambda} = \boldsymbol{0} \in I/2$, hence it might be advantageous to think of $\hat{K}'_m(I/2)$ as a region approximately centred around $\boldsymbol{\mu}$. In Section 3 we build upon this interpretation.

3. Extended empirical saddlepoint approximation

The aim of this work is to use the flexibility of the empirical saddlepoint approximation to estimate densities for which the normal approximation is poor. The asymptotic results of Feuerverger (1989) suggest that the saddlepoint approximation should perform reasonably well in the central part of the distribution, while its accuracy decreases in the tails. More importantly, as stated in Section 2.1, the empirical saddlepoint equation (6) has a solution only if $\boldsymbol{s}$ lies inside the convex hull of the simulated data, so the resulting empirical saddlepoint density is not defined outside this subset of $\mathbb{R}^d$. This is problematic in the context of SL because, whether we wish to estimate the unknown parameters by Maximum Likelihood or MCMC, we cannot generally expect $\boldsymbol{s}^0$ to fall inside the convex hull of the simulated statistics in early iterations. In addition, if the model of interest is unable to generate summary statistics that are close to the observed ones, its inadequacy should ideally be quantified by a low, rather than an undefined, value of the synthetic likelihood. Hence, we need a remedy that allows us to solve (6) for any $\boldsymbol{s} = \boldsymbol{s}^0$.

To motivate our solution, notice that solving (6) is equivalent to minimizing

$$\{\hat{K}_m(\boldsymbol{\lambda}) - \boldsymbol{\lambda}^T \boldsymbol{s}\}^2,$$

which would be guaranteed to have a unique minimum, if strong convexity held. That is, if

$$\exists\, \epsilon \in \mathbb{R}^+ \text{ such that } \boldsymbol{z}^T \hat{K}_m''(\boldsymbol{\lambda})\boldsymbol{z} > \epsilon ||\boldsymbol{z}||^2, \quad \forall\, \boldsymbol{\lambda}, \boldsymbol{z} \in \mathbb{R}^d \text{ such that } ||\boldsymbol{z}|| > 0, \tag{7}$$

then (6) could be solved for any $\boldsymbol{s}$. Unfortunately, the following proposition states that this in not the case.

Proposition 1. $\hat{K}_m(\boldsymbol{\lambda})$ *is strictly, but not strongly, convex.*

Proof. See Appendix A.1. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$

To address this problem, we propose to tilt $\hat{K}(\boldsymbol{\lambda})$ using a strongly convex function. In particular, we propose to use a modified estimator

$$\hat{K}_m(\boldsymbol{\lambda}, \gamma, \boldsymbol{s}) = g(\boldsymbol{s}, \gamma)\hat{K}_m(\boldsymbol{\lambda}) + \{1 - g(\boldsymbol{s}, \gamma)\}\hat{G}_m(\boldsymbol{\lambda}), \tag{8}$$

where $\hat{G}_m(\boldsymbol{\lambda})$ is a strongly convex function, while $g(\boldsymbol{s}, \gamma)$ is a function of $\boldsymbol{s}$, parametrized by γ, which determines the mix between the two functions. Furthermore, we require

$$g(\boldsymbol{s}, \gamma) : \mathbb{R}^d \to [0, 1],$$
$$g(\boldsymbol{s} = \hat{\boldsymbol{\mu}}, \gamma) = 1, \qquad g(\boldsymbol{s} \neq \hat{\boldsymbol{\mu}}, \gamma) < 1,$$
$$\lim_{||\boldsymbol{s} - \hat{\boldsymbol{\mu}}|| \to \infty} g(\boldsymbol{s}, \gamma) = 0. \tag{9}$$

If estimator (8) is adopted and conditions (9) are met, then the saddlepoint equation (6) can be solved for any $\boldsymbol{s}$. In fact, at $\boldsymbol{s} = \hat{\boldsymbol{\mu}}$, we have that $\hat{K}_m(\boldsymbol{\lambda}, \gamma, \boldsymbol{s}) = \hat{K}_m(\boldsymbol{\lambda})$ and the solution of (6) is $\hat{\boldsymbol{\lambda}} = \boldsymbol{0}$. If $\boldsymbol{s} \neq \hat{\boldsymbol{\mu}}$, then $\hat{K}_m(\boldsymbol{\lambda}, \gamma, \boldsymbol{s})$ is strongly convex, because the sum of a strictly and a strongly function is itself strongly convex. This can be verified using the fact that the smallest eigenvalue of the Hessian matrix of a strongly convex function is bounded from below by a positive constant, together with the fact that the smallest eigenvalue of the sum of two positive definite matrices is larger than the sum of their smallest eigenvalues.

A natural choice for $\hat{G}_m(\boldsymbol{\lambda})$ is the following parametric estimator of $K(\boldsymbol{\lambda})$

$$\hat{G}_m(\boldsymbol{\lambda}) = \boldsymbol{\lambda}^T \hat{\boldsymbol{\mu}} + \frac{1}{2}\boldsymbol{\lambda}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\lambda}, \tag{10}$$

which is unbiased and consistent for multivariate normal random variables. This leads to

$$\hat{K}_m(\boldsymbol{\lambda}, \gamma, \boldsymbol{s}) = \boldsymbol{\lambda}^T \hat{\boldsymbol{\mu}} + \frac{\eta}{2}\boldsymbol{\lambda}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\lambda}$$
$$+ g(\boldsymbol{s}, \gamma)\left\{\frac{1}{3!}\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\frac{\partial^3 \hat{K}_m}{\partial \lambda_i \partial \lambda_j \partial \lambda_k}\bigg|_{\boldsymbol{\lambda}=\boldsymbol{0}}\lambda_i \lambda_j \lambda_k + \cdots\right\},$$

where $\eta = 1 - g(\boldsymbol{s}, \gamma)/(m-1)$ appears because here $\hat{\boldsymbol{\Sigma}}$ is the standard unbiased covariance estimator, while $\hat{K}_m''(\boldsymbol{\lambda} = \boldsymbol{0}) = m\hat{\boldsymbol{\Sigma}}/(m-1)$. Similarly, evaluating higher derivatives of $\hat{K}_m(\boldsymbol{\lambda})$ at $\boldsymbol{\lambda} = \boldsymbol{0}$ produces consistent, but biased, estimators of the corresponding cumulants. Unbiased cumulant estimators are the k-statistics (McCullagh, 1987).

Our solution is related to that of Wang (1992), who modified the truncated estimator of Easton and Ronchetti (1986), and to the proposal of Bartolucci (2007), in the context of Empirical Likelihood (Owen, 2001). We refer to the density obtained by using estimator (8) within (5) as the Extended Empirical Saddlepoint approximation (EES). In Section 3.1 we propose a particular form for $g(\boldsymbol{s}, \gamma)$.

3.1. Choosing and tuning the mixture function $g(\boldsymbol{s}, \gamma)$

In Appendix A.2 we derive the MSEs of estimators (4) and (10), under normality of $\boldsymbol{S}$. We then base our choice of $g(\boldsymbol{s}, \gamma)$ on the relative MSE performance of the two estimators. In particular, we choose

$$
\begin{aligned}
g(\boldsymbol{s}, \gamma) &= \left[\frac{(\boldsymbol{s} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{s} - \hat{\boldsymbol{\mu}}) \left\{ 1 + \frac{1}{2} (\boldsymbol{s} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{s} - \hat{\boldsymbol{\mu}}) \right\} + 1}{\exp\left\{ (\boldsymbol{s} - \hat{\boldsymbol{\mu}})^T \hat{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{s} - \hat{\boldsymbol{\mu}}) \right\}} \right]^{\gamma} \\
&\approx \left[\frac{\mathrm{MSE}\{\hat{G}_m(\boldsymbol{\lambda})\} + 1}{\mathrm{MSE}\{\hat{K}_m(\boldsymbol{\lambda})\} + 1} \right]^{\gamma},
\end{aligned}
\tag{11}
$$

where $\gamma > 0$ is a tuning parameter, which determines the rate at which $g(\boldsymbol{s}, \gamma)$ varies from 1 to 0, as the distance between $\boldsymbol{s}$ and $\hat{\boldsymbol{\mu}}$ increases. Apart from fulfilling requirement (9), function (11) has the desirable property of being invariant under linear transformations. More precisely, if $\boldsymbol{z} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{s}$ and $\boldsymbol{Z}_i = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{S}_i$, for $i = 1, \ldots, m$, then $g^{\boldsymbol{z}}(\boldsymbol{z}, \gamma) = g(\boldsymbol{s}, \gamma)$. Using this fact, it is simple to show that EES is equivariant under such transformations, that is $\log \hat{p}_m^{\boldsymbol{z}}(\boldsymbol{z}, \gamma) = \log \hat{p}_m(\boldsymbol{s}, \gamma) - \log \det(\boldsymbol{B})$. In practice, this allows us to normalize $\boldsymbol{s}$ and $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_m$ before fitting, which generally enhances numerical stability. This is achieved by setting $\boldsymbol{a} = -\hat{\boldsymbol{\mu}}$ and $\boldsymbol{B} = \mathrm{diag}(\hat{\boldsymbol{\Sigma}})^{-1/2}$, so that $\log \det(\boldsymbol{B}) = -\sum_{i=1}^{d} \log \hat{\Sigma}_{ii}/2$.

Our choice (11) has two main shortcomings: it is based on a normality assumption for $\boldsymbol{S}$ and, most importantly, it does not take the sample size m into account. In regard to the first issue: using higher moments of the simulated statistics to determine (11) might be attractive, but our experience suggests that this would result in very unstable estimates. The second problem can be addressed by appropriately selecting the tuning parameter γ. Its value is critical for the performance of our method, and at first sight it not clear on what principle this choice should be based. However, saddlepoint approximations are exact for Gaussian densities (Butler, 2007), hence γ is fundamentally a complexity-controlling parameter, which determines the balance between two density estimators: the empirical saddlepoint, which is characterized by higher flexibility and variance, and the normal distribution, which generally has higher bias, but lower variance. Hence, we propose to select γ by k-fold cross-validation, as detailed in the Algorithm 2.

In Appendix B.2 we show that, as m and $l \to \infty$, Algorithm 2 consistently selects the value of γ which minimizes the Kullback-Leibler divergence between $\bar{p}(\boldsymbol{s}, \gamma)$ and $p(\boldsymbol{s})$. The Gaussian case is recovered as $\gamma \to \infty$.

Algorithm 2 Cross-validation with nested normalization

1: Create a grid of r possible values, $\gamma_1, \ldots, \gamma_r$, for the tuning parameter.
2: Simulate m random d-dimensional vectors $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_m$ from the true density $p(\boldsymbol{s})$ and divide them into k folds. For simplicity, assume that m is a multiple of k. Indicate with $\bar{\boldsymbol{S}}_t$ the vectors in the t-th fold, and with $\bar{\boldsymbol{S}}_{-t}$ the remaining $r = m(1 - 1/k)$ vectors. Let $\hat{p}_r^{-t}(\boldsymbol{s}, \gamma)$ be the EES density based on the vectors in $\bar{\boldsymbol{S}}_{-t}$.
3: For $i = 1, \ldots, r$

  For $t = 1, \ldots, k$

   · Estimate the normalizing constant of $\hat{p}_r^{-t}(\boldsymbol{s}, \gamma_i)$ by importance sampling, that is

$$\hat{z}_r^{-t}(\gamma_i) = \frac{1}{l} \sum_{j=1}^{l} \frac{\hat{p}_r^{-t}(\boldsymbol{S}_j^*, \gamma_i)}{q(\boldsymbol{S}_j^*)}, \quad \boldsymbol{S}_j^* \sim q(\boldsymbol{s}), \quad \text{for } j = 1, \ldots, l.$$

   A reasonably efficient d-dimensional importance density $q(\boldsymbol{s})$ can be obtained by constructing a multivariate normal density with mean and covariance equal to the sample mean and covariance of $\bar{\boldsymbol{S}}_{-t}$. Notice that (9) and (10) assure the boundedness of the importance weights, under this choice of $q(\boldsymbol{s})$.

   · Using the normalized EES density,

$$\bar{p}_r^{-t}(\boldsymbol{s}, \gamma_i) = \frac{\hat{p}_r^{-t}(\boldsymbol{s}, \gamma_i)}{\hat{z}_r^{-t}(\gamma_i)},$$

   evaluate the negative log-likelihood $-\log \bar{p}_r^{-t}(\bar{\boldsymbol{S}}_t, \gamma_i)$.
4: Select the value γ_i that minimizes the negative validation log-likelihood, averaged across the k folds.

---

4. Use within synthetic likelihood

We now describe how the proposed density estimator can be used within the context of SL, hence we restore all dependencies on the model parameters, $\boldsymbol{\theta}$. We obtain an approximate estimate of the parameters, $\boldsymbol{\theta}_I$, by maximizing the synthetic likelihood based on the Gaussian approximation. This step allows us to move toward a more likely area of the parameter space, and typically it does not add much to the computational cost, as the Gaussian density is cheaper to evaluate and can be estimated using fewer simulated statistics. Then, γ can be selected using Algorithm 2, with $p(\boldsymbol{s}) = p(\boldsymbol{s}|\boldsymbol{\theta}_I)$. Given γ, pointwise estimates of the synthetic likelihood can be based on the new density estimator by using a procedure analogous to Algorithm 1, which we describe in Appendix C.1.

In terms of computational effort, if we assume that m, the number of summary statistics simulated from $p(\boldsymbol{s}|\boldsymbol{\theta})$, is much larger than d, then the cost of evaluating the Gaussian synthetic likelihood is $O(md^2)$, which is the cost of obtaining $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$. Calculating $\hat{K}_m''(\boldsymbol{\lambda})$ has the same complexity, but solving the empirical saddlepoint equation (6) numerically implies that $\hat{K}_m''(\boldsymbol{\lambda})$ will be evaluated at several values of $\boldsymbol{\lambda}$. The proposal described in Section 3 assures that the underlying root finding problem is strongly convex, hence few iterations of a Newton-Raphson algorithm are generally sufficient to solve (6) with high accuracy. The computational cost of a synthetic likelihood estimate is then $O(lmd^2)$, if the normalizing constant is estimated using l importance samples. In prac-

tice, we have not yet encountered an example where the normalizing constant strongly depended on $\boldsymbol{\theta}$. However, the normalizing constant often varies significantly with γ. Hence, in the examples presented in this paper, we estimate the normalizing constant when selecting γ using Algorithm 2, but we use the unnormalized EES density during parameter estimation.

Before testing the ESS-based version of SL on the examples, we now prove that, under the conditions to be specified shortly, maximizing the synthetic likelihood leads to consistent parameter estimators. Here we denote the Gaussian-based synthetic likelihood with $\hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta})$ and its EES-based version with $\hat{p}_S(\boldsymbol{s}^0|\boldsymbol{\theta})$. We firstly consider the Gaussian case and we prove identifiability, which means that the (scaled) synthetic likelihood converges to a function which is maximized at the true parameter vector, $\boldsymbol{\theta}_0$. This is guaranteed under the following assumptions.

Assumption 1. *The summary statistics depend on a set of underlying observations $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$, and have mean and covariance matrix*

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}^n = \mathbb{E}(\boldsymbol{S}_n \,|\, \boldsymbol{\theta}), \quad \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^n = \mathbb{E}\big\{(\boldsymbol{S}_n - \boldsymbol{\mu}_{\boldsymbol{\theta}}^n)(\boldsymbol{S}_n - \boldsymbol{\mu}_{\boldsymbol{\theta}}^n)^T \,|\, \boldsymbol{\theta}\big\},$$

where $\boldsymbol{S}_n = S(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$. In addition there exists $\delta > 0$ such that, for any $\boldsymbol{\theta}$, we have

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n \to \boldsymbol{\mu}_{\boldsymbol{\theta}} \quad and \quad n^\delta \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n \to \boldsymbol{\Sigma}_{\boldsymbol{\theta}},$$

in probability, as m and $n \to \infty$.

Assumption 2. *Let $_*\psi_{\boldsymbol{\theta}}$ and $^*\psi_{\boldsymbol{\theta}}$ be, respectively, the smallest and the largest eigenvalues of the asymptotic (scaled) covariance matrix, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$. There exists two positive constants, $_*\psi$ and $^*\psi$, such that $_*\psi_{\boldsymbol{\theta}} > _*\psi$ and $^*\psi_{\boldsymbol{\theta}} < ^*\psi$ for any $\boldsymbol{\theta}$.*

Assumption 3. *$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mu(\boldsymbol{\theta})$ is one to one.*

Theorem 1. *If Assumptions 1 to 3 hold, as m and $n \to \infty$ the scaled synthetic log-likelihood, $n^{-\delta} \log \hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta})$, is asymptotically proportional to a function, $f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$, which is maximal at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.*

Proof. *See Appendix A.3.*

Here Assumption 1 guarantees pointwise convergence to $f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$, while Assumptions 2 and 3 assure identifiability. The fact that $f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$ is maximal at the true parameter is not itself sufficient to assure weak consistency, which is instead guaranteed under the additional condition that the convergence of the Gaussian synthetic likelihood is uniform (Van der Vaart, 2000). To assure this, we make the following assumptions.

Assumption 4. *The parameter space, $\boldsymbol{\Theta} \subset \mathbb{R}^p$, is compact and convex.*

Assumption 5. *The derivatives of $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n$ and $n^\delta \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n$ are continuous and dominated by two $O_p(1)$ positive random sequences, $a_{n,m}$ and $b_{n,m}$. More precisely*

$$\left\|\frac{\partial \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n}{\partial \theta_k}\right\|_2 \leq a_{n,m}, \quad and \quad \left\|\frac{\partial n^\delta \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n}{\partial \theta_k}\right\|_2 \leq b_{n,m},$$

for $k = 1, \ldots, q$ and for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Assumption 6. *Let $_*\hat{\psi}^n_{\boldsymbol{\theta}}$ and $^*\hat{\psi}^n_{\boldsymbol{\theta}}$ be, respectively, the smallest and the largest eigenvalue of $n^\delta \hat{\boldsymbol{\Sigma}}^n_{\boldsymbol{\theta}}$ and assume that there exist two $O_p(1)$ positive random sequences, $c_{n,m}$ and $u_{n,m}$, such that*

$$\left(_*\hat{\psi}^n_{\boldsymbol{\theta}}\right)^{-1} \leq c_{n,m}, \quad and \quad ^*\hat{\psi}^n_{\boldsymbol{\theta}} \leq u_{n,m},$$

for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Assumption 7. *The derivatives of the asymptotic mean vector, $\boldsymbol{\mu_\theta}$, and (scaled) covariance matrix, $\boldsymbol{\Sigma_\theta}$, are bounded. In particular, there exist two positive constants, $M_{\boldsymbol{\mu}}$ and $M_{\boldsymbol{\Sigma}}$, such that*

$$\left\|\frac{\partial \boldsymbol{\mu_\theta}}{\partial \theta_k}\right\|_2 \leq M_{\boldsymbol{\mu}}, \quad and \quad \left\|\frac{\partial \boldsymbol{\Sigma_\theta}}{\partial \theta_k}\right\|_2 \leq M_{\boldsymbol{\Sigma}},$$

for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Newey (1991) shows that pointwise convergence in probability, proved as part of Theorem 1, implies uniform convergence as long as: Assumption 4 holds; the derivatives of $n^{-\delta} \log \hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta})$ are continuous and dominated by an $O_p(1)$ random sequence; $f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$ is equicontinuous. Hence, in the Gaussian case, proving consistency requires only assuring that the last two requirements are met.

Theorem 2. *Let $\hat{\boldsymbol{\theta}}$ be the maximizer of the Gaussian synthetic likelihood. If Assumptions 1 to 7 hold, then $\hat{\boldsymbol{\theta}}$ converges in probability to $\boldsymbol{\theta}_0$, as m and $n \to \infty$.*

Proof. *See Appendix A.4.*

We now consider the EES-based synthetic likelihood, and we focus on the un-normalized density estimator, which is cheaper to compute in practice. To prove identifiability we require the two following conditions to hold, in addition to Assumptions 1, 2 and 3.

Assumption 8. *For every n, the moment generating function of $\boldsymbol{S}_n$ exists for $\boldsymbol{\lambda} \in I$, where I is a non-vanishing subset of $\mathbb{R}^d$ containing the origin.*

Assumption 9. *Let $\hat{\gamma}^n_{\boldsymbol{\theta}_I}$ be the chosen tuning parameter, corresponding to simulation effort m and sample size n. As m and $n \to \infty$, there exists a constant $c > 0$ such that*

$$Prob(\hat{\gamma}^n_{\boldsymbol{\theta}_I} < c) \to 0,$$

for any initialization $\boldsymbol{\theta}_I$.

Theorem 3. *If Assumptions 1, 2, 3, 8 and 9 hold, as m and $n \to \infty$ the scaled synthetic log-likelihood, $n^{-\delta} \log \hat{p}_S(\boldsymbol{s}^0|\boldsymbol{\theta})$, is asymptotically proportional to a function, $f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$, which is maximal at $\boldsymbol{\theta}_0$.*

Proof. *See Appendix A.5.*

Notice that the asymptotic function, $f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$, mentioned in Theorems 1 and 3, is the same under either density estimator. As in the Gaussian case, weak consistency is guaranteed under identifiability and uniform convergence to $f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$.

Given Assumption 4, and the fact that the equicontinuity of $f_{\theta_0}(\theta)$ has already been proven in the proof of Theorem 2, uniform convergence is guaranteed as long as the derivatives of $n^{-\delta} \log \hat{p}_S(s^0|\theta)$ are continuous and dominated by an $O_p(1)$ sequence. In the Gaussian case this was assured under assumptions on the derivatives of $\hat{\mu}_{\theta}^n$ and $n^{\delta}\hat{\Sigma}_{\theta}^n$, and on the eigenvalues of the latter. Given the complexity of the EES density, under this density estimator we prefer to impose conditions directly on $n^{-\delta} \log \hat{p}_S(s^0|\theta)$, rather than on $K(\lambda)$, $K'(\lambda)$ and $K''(\lambda)$.

Assumption 10. *The derivatives of the synthetic log-likelihood based on the EES density are continuous and dominated by an $O_p(1)$ random sequence, $v_{n,m}$, that is*

$$\left| \frac{\partial n^{-\delta} \log \hat{p}_S(s^0|\theta)}{\partial \theta_k} \right| \leq v_{n,m},$$

for $k = 1, \ldots, q$ and for any $\theta \in \Theta$.

Theorem 4. *Let $\hat{\theta}$ be the maximizer of the EES-based synthetic likelihood. If Assumptions 1, 2, 3, 4, 7, 8, 9 and 10 hold, then $\hat{\theta}$ converges weakly to θ_0, as m and $n \to \infty$.*

Proof. *Theorem 3 assures pointwise convergence and identifiability under Assumptions 1, 2, 3, 8 and 9. Pointwise converge, together with Assumptions 4, 7 and 10 guarantee uniform convergence in probability (Newey, 1991). Uniform convergence and identifiability are sufficient conditions for weak consistency (Van der Vaart, 2000).*

In the examples, we maximize the synthetic likelihood using a special case of the Iterated Filtering (IF) procedure, firstly proposed by Ionides et al. (2006). The convergence properties of this algorithm have been studied in detail by Ionides et al. (2011) in the context of state space models, while Doucet et al. (2013) explicitly pointed out that IF can be used as a general likelihood optimizer. While Doucet et al. (2013) considered cases where the likelihood can be evaluated exactly, we have verified empirically that the algorithm works well also when the likelihood is estimated with Monte Carlo error. We preferred using IF over stochastic gradient methods based of finite differences (see e.g. Spall (2005)), because at each iteration IF updates the most recent parameter estimate using a convex combination of perturbed parameter vectors. In our experience this convex update makes unlikely that the optimizer will suddenly diverge when the update is performed. Appendix C.2 details the basic steps of this procedure.

5. Simple recruitment, boom and bust model

Figure 1a shows two trajectories simulated from model (1), using parameters $r = 0.4$, $\kappa = 50$, $\alpha = 0.09$ and $\beta = 0.05$. To compare ABC with the Gaussian and EES-based version of SL, we simulate 100 pseudo-observed datasets of length $T = 300$ using the above parameters. Given that we are not interested in estimating N_0, we discards the first 50 times steps to lose any transient behaviour
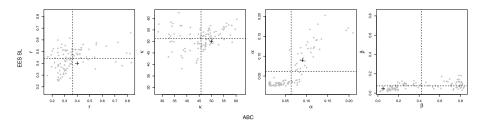
FIG 2. *MLE estimates obtained using EES SL versus MAP estimates produced by ABC, under model (1). The dashed lines and the black crosses indicate, respectively, the mean estimates under each method and the true parameter values.*

from the system. We use the remaining 250 steps of each trajectory to compute the following summary statistics: mean and smallest population, number of times the population consists of one or less individuals, number of population peaks and square-root of the minimal time gap between two consecutive peaks (a peak is occurring at time t if $N_{t+1} - N_t \leq -30$). Under ABC, we obtain MAP estimates of the model parameters using the approach of Rubio et al. (2013). This consists of sampling the approximate posterior, and then maximizing a kernel density estimate of it. We perform the sampling step using the MCMC approach of Marjoram et al. (2003), and we maximize the approximate posterior using the mean shift algorithm (Fukunaga and Hostetler, 1975). The ABC tolerance was chosen using the approach of Wegmann et al. (2009), which involves simulating a large number of parameter vectors from the prior, then simulating a statistics vector from the model using each of them and selecting ϵ so that only a small percentage of these are accepted. We use 10^6 simulations for tuning and 10^{-3} as target acceptance rate. The scaling matrix for ABC is a diagonal matrix, $\boldsymbol{Q}$, where $\boldsymbol{Q}_{ii}$ is the reciprocal of the sample standard deviation of the i-th summary statistic, calculated using all the 10^6 simulated statistics vectors. We use the uniform priors $r \in (0,1)$, $\kappa \in (10, 80)$, $\alpha \in (0,1)$ and $\beta \in (0,1)$, so that MLE and MAP estimates are equivalent. For SL we use $m = 5 \times 10^3$ and estimate $\gamma = 4 \times 10^{-4}$ using Algorithm 2, with $l = 10^4$. While Appendix C.4 gives further details about the simulation setting, it is important to point out here that we use the same number of simulations from the model under all methods.

Figure 2 compares the estimates obtained using ABC and EES-based SL, while Table 1 reports the true parameters, together with the means and RMSEs for all three methods. ABC struggles to identify the arrival rate, β, and it often grossly overestimates r and underestimates κ, so that its RMSE performances is substantially worse overall than that of either SL method. Comparing the synthetic likelihood methods to each other, EES-SL has lower RMSE than Gaussian SL for all parameters. The mean for β is also substantially closer to the true parameter for EES-SL, as expected given the shape of the distribution of the most relevant statistic, shown in Figure 1c. This very simple example clearly illustrates that EES-SL offers non-negligible benefits when important statistics

Param.	Truth	ABC	Gaus. SL	EES SL
r	0.4	0.36(0.15)	0.45(0.11)	0.44(0.097)
κ	50	45.7(8.5)	51.7(4.8)	51.4(4.5)
α	0.09	0.063(0.05)	0.062(0.057)	0.061(0.054)
β	0.05	0.424(0.467)	0.105(0.073)	0.075(0.044)

TABLE 1

True parameters, means and RMSEs (in parentheses) of the estimates using ABC and the two version of SL, for model (1). For each parameter, the lowest RMSE is underlined.
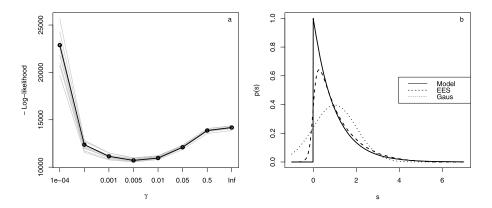

FIG 3. *Shifted exponential model. a: Curves from 10-fold cross-validation, the black line is their average. b: True Exp(β) density (black), EES (dashed) and normal (dotted) approximation.*

are highly non-Gaussian.

6. Multivariate shifted exponential distribution

Here we consider a toy example, whose purpose is illustrating how the performance of the Gaussian and EES versions of SL compares with that of tolerance-based ABC algorithms, as the dimensionality of the model increases. In particular, let $\boldsymbol{S}$ be a d-dimensional random vector, where each marginal follows a shifted exponential distribution

$$S_k \sim \theta_k + \mathrm{Exp}(\beta), \quad \text{for } k = 1, \dots, d. \tag{12}$$

Here we choose $\beta = 0.5$ and $\theta_1 = \cdots = \theta_d = 0$. The plot in Figure 3a contains the results of a 10-fold cross-validation run, obtained using $d = 10$, $l = 10^3$ and $m = 10^4$. The cross-validation curve is minimized by $\gamma = 5 \times 10^{-3}$, and the plot in Figure 3b shows the true and approximate marginal densities of one component S_k. The EES approximation to the marginal, obtained by marginalizing the d-dimensional fit, is clearly more accurate than a normal density.

To demonstrate the usefulness of EES in the context of SL, we use it to estimate the shifts $\theta_1, \dots, \theta_d$, all of which are equal to 0. In particular, we

simulate a single vector of observed statistics, $\boldsymbol{s}^0$, from (12), and we maximize the resulting synthetic likelihood, using either a Gaussian density or EES. Given that SL and ABC algorithms are often motivated as approximations to full likelihood inference, all the MSEs reported in the remainder of this section quantify deviations from the full MLE, which is $\boldsymbol{s}^0$, not from the true parameters. Hence, the bias of Gaussian SL estimates is $1/\beta$. By averaging the squared errors across the 10 dimensions, we obtain MSEs equal to 3.8 and 0.56, using the normal and the EES approximation respectively. In an analogous 20-dimensional run, using $m = 5 \times 10^4$, the MSE was reduced from 4.1 to 1.26. P-values from t-test for differences in log-absolute errors were around 10^{-6} in both runs.

It is possible to derive analytically how a tolerance-based ABC approximation would perform under this model. The details are reported in Appendix C.3. Assume that the likelihood $p(\boldsymbol{s}^0|\boldsymbol{\theta})$ is approximated by

$$p_\epsilon(\boldsymbol{s}^0|\boldsymbol{\theta}) = p(||\boldsymbol{s}_0 - \boldsymbol{s}||_\infty < \epsilon|\boldsymbol{\theta}) = \prod_{k=1}^{d} p_\epsilon(s_k^0|\theta_k),$$

where

$$p_\epsilon(s_k^0|\theta_k) = p(|s_k - s_k^0| < \epsilon|\theta_k) = \int_0^\infty \mathbb{I}(|u - s_k^0| < \epsilon)p(u|\theta_k)du,$$

for $k = 1, \ldots, d$. Here $\epsilon > 0$ is the ABC tolerance and $\mathbb{I}()$ is an indicator function. Given that we are interested in deviations from the full MLE, we can impose $\boldsymbol{s}^0 = \boldsymbol{0}$ without loss of generality. If we use independent uniform priors on $[\psi, 0]$ for each parameter, where $\psi < -\epsilon$, the posterior mode is at $\theta_k = -\epsilon$, for $k = 1, \ldots, d$. Hence, the MSE corresponding to the MAP estimate is equal to ϵ^2. This implies that, to achieve MSEs equal to those of EES, ϵ would need to be set to $\sqrt{0.56}$ and $\sqrt{1.26}$, respectively in the 10 and 20-dimensional setting. The corresponding acceptance probabilities, obtained by simulating statistic vectors using parameters $\boldsymbol{\theta}$ fixed to the MAP, are of order 10^{-3} and 10^{-4}. In 40 dimensions, obtaining an MSE equal to 2 would lead to an acceptance ratio at the MAP of order 10^{-5}. Notice that these are upper bounds, because the acceptance probability is maximal at the MAP.

This analysis suggests that, in order to match the MSE achieved by SL, the computational budget of an ABC algorithm would need to be increased rapidly as the number of dimensions grows. Further, the relation between ϵ and the MSE is generally not known in practice. A popular approach, which we adopted in Section 5, is to simulate many parameter vectors from the prior, use each of them to simulate a statistics vector from the model and select ϵ so that a small percentage of the simulated parameters and statistics are accepted. To quantify how the computational cost of this tuning phase depends on the prior, we reverse this process and assume that the values of ϵ are as given above. In 10 dimensions, ψ would need be in $[-2.1, -\epsilon]$, in order to achieve an acceptance probability of order 10^{-4}, while in 20 dimension, $\psi \in [-1.9, -\epsilon]$ leads an acceptance probability of order 10^{-5}. Hence, to obtain only just tolerable acceptance rates during the tuning phase, very accurate prior information must be available,

especially in high dimensions. In an applied setting, prior information is often rather vague, hence ϵ needs to be tuned using more sophisticated approaches, such as the sequential algorithm of Toni et al. (2009). While such methods can alleviate the effort needed to select ϵ, performing extensive ABC tuning runs is still onerous when working with computationally intensive models, such as the one described in the next section.

7. Formind forest model

7.1. The model

To test our proposal in a realistic setting, we consider Formind, an individual-based model describing the main natural processes driving forests dynamics. Here we describe its basic features, while we refer to Dislich et al. (2009) and to Fischer et al. (2016) for detailed descriptions of the model and of the scientific questions it can be used to address.

The model describes the growth and population dynamics of tree individuals in a simulation area that is divided in 20×20m patches, with individual trees being assigned explicitly to one patch. Tree species with similar characteristics are grouped into Plant Functional Types (PFTs). A constant input of seeds deposits on average s_j seeds of the j-th PFT per hectare per year. The main factor determining both seed establishment and growth is the light climate in the patch. For example, pioneer types will establish only in patches relatively free of overshadowing trees, while late successional trees are able to grow below a dense canopy. Trees are subject to a baseline mortality rates m_j, which is specific to each PFT.

In the context of Formind, the need for approximate simulation-based methods comes from the complexity of the model. Indeed, Formind was developed with a focus on ecological plausibility, rather than statistical tractability, and most of its submodels describe highly non-linear biological processes, containing one or more sources of randomness. Most importantly, the raw output of Formind is the collection of all the characteristics of individual trees in the simulations area, which obviously do not correspond to individuals present in the actual survey data. Hence, it is necessary to work with summary statistics.

Formind is computationally intensive even when few PFTs are included and, given initial conditions and parameters, the simulated forest needs to be run for several hundred years, before the distribution of the summary statistics reaches equilibrium. This means that, from a practical point of view, it is critical to avoid lengthy tuning runs, such as those needed to select the tolerance ϵ in ABC methods.

7.2. Simulation results

We consider two PFTs, pioneer and late successional, and we reduce the model output to 6 summary statistics. In particular, to verify whether then new density

Param.	Truth	Gaus. SL	EES SL	P-value
μ_{pio}	0.05	0.047 (0.014)	0.054 (0.007)	0.002
μ_{suc}	0.005	0.0093 (0.0065)	0.0061 (0.0016)	0.003
s_{pio}	80	108.4 (41.1)	91.6 (26.2)	0.07
s_{suc}	20	31.6 (15.7)	23.2 (4.7)	0.003

TABLE 2

Formind model: true parameters, means and RMSEs (in parentheses) of the estimates using the normal and the EES estimators. P-values for differences in log-absolute errors have been calculated using t-tests.

estimator can deal with large departures from normality, we used the following transformed statistics

$$S_{jk} = \alpha_{jk}^{\frac{C_{jk} - \psi_{jk}}{\sigma_{jk}}}, \ \ \text{for } j \in \{1, 2\}, \ k \in \{1, 2, 3\},$$

where C_{jk} is the number of trees of the j-th PFT falling in the k-th diameter class, while α_{jk}, ψ_{jk}, and σ_{jk} are constants, whose values are reported in Appendix C.5. The diameter categories used for each PFT correspond to trees with small, medium or large diameters.

We simulated 24 datasets from the model and estimated the baseline mortality rates and seed input intensities of the two PFTs by maximizing the synthetic likelihood, using both the normal and the EES approximations. In both cases we used $m = 10^4$ simulated summary statistics and, under EES, γ was fixed to 5.5×10^{-3}, chosen using Algorithm 2 with $l = 10^3$. Table 2 reports the true parameters, together with the means and RMSEs of the estimates, from the normal or the EES approximations. See Appendix C.5 for more details about the optimization setting.

Using the EES, rather than the normal approximation, leads to lower MSEs for all model parameters. The plots in Figure 4 compare the marginal distributions of the summary statistics, simulated from the model using the true parameter values, with those obtained by simulating random vectors from EES, fitted to the simulated statistics using the same values of γ and m used during the optimization. To simulate random vectors from EES we used importance sampling with the same Gaussian importance density used in Algorithm 2, and we then obtained an un-weighted sample by re-sampling the resulting 2×10^5 importance samples. EES gives a good fit to the marginal distributions of the summary statistics, all of which are far from normal.

8. Conclusions

In this work we have relaxed the normality assumption, which characterized the original formulation of SL, by proposing a novel, more flexible, density estimator. As the examples show, EES scales well with the number of summary statistics used and it is able to model densities for which a normal approximation is clearly inadequate. This in turn can lead to better accuracy in parameter estimation.
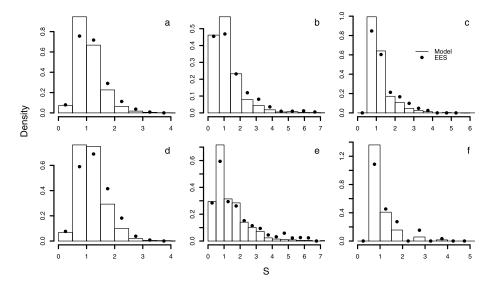
FIG 4. *Marginal distributions of summary statistics corresponding to small, medium and large pioneers (a, b, c) and successionals (d, e, f), in the Formind model.*

Importantly, using EES rather than a Gaussian density, does not add much to the tuning requirements of SL. In fact, the only parameter of EES, γ, can be selected using standard statistical tools, such as cross-validation. In the context of SL, and of approximate methods in general, having little tuning requirements is an important feature, since it allows practitioners to focus on identifying informative summary statistics, rather than on other aspects of the inferential procedure.

We have shown that, if fairly general conditions on the distribution of the summary statistics and on the underlying model hold, maximizing the synthetic likelihood function leads to consistent parameter estimators, under either EES or a Gaussian density estimator. Given the generality of the conditions assumed here, we have treated EES as a non-parametric density estimator, as suggested by Feuerverger (1989). However, the use of empirical saddlepoint approximations has previously been considered for particular classes of statistics, such as M-estimators (Monti and Ronchetti, 1993; Ronchetti and Welsh, 1994) and L-statistics (Easton and Ronchetti, 1986). Hence, it would be interesting to verify whether making additional assumptions on the summary statistics would allow us to assess the asymptotic efficiency of the parameter estimates produced by the EES-based version of SL or, as suggested by a reviewer, to prove the consistency of $\hat{\boldsymbol{\beta}}$ by exploiting pre-existing results concerning the relative error properties of empirical saddlepoint approximations.

From a practical point of view, the computational efficiency of SL is of critical importance. When the ESS density is normalised by importance sampling, the saddlepoint equation has to be solved for each importance sample, $\boldsymbol{S}_j^*$. Given

that samples that are close to each other should lead to similar solutions to (6), it might be possible to devise an efficient scheme that reduces the number of Newton steps needed to solve (6) by using neighbouring solutions for initialization. Another possibility would be adapting the methods of Gutmann and Corander (2016), Wilkinson (2014) and Meeds and Welling (2014), who proposed using Gaussian Processes to increase the computational efficiency of SL and ABC methods. The first two proposals, being based on pointwise likelihood estimates, could be used in conjunction with EES. Meeds and Welling (2014) model only the first two moments of the simulated statistics, hence it is not clear whether their approach could be modified to take higher moments into account, as the new density estimator does.

Appendix A: Main theoretical results

A.1. Proof of Proposition 1

Define

$$w_i = \frac{e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i}}{\sum_{i=1}^m e^{\boldsymbol{\lambda}^T \boldsymbol{s}_i}}, \quad \bar{\boldsymbol{s}} = \hat{K}'(\boldsymbol{\lambda}) = \frac{\sum_{i=1}^m w_i \boldsymbol{s}_i}{\sum_{i=1}^m w_i}, \quad \mathrm{i} = 1, \ldots, m, \qquad (13)$$

and notice that $\hat{K}''(\boldsymbol{\lambda})$ is positive semi-definite

$$
\begin{aligned}
\boldsymbol{z}^T \hat{K}''(\boldsymbol{\lambda}) \boldsymbol{z} &= \boldsymbol{z}^T \sum_{i=1}^m w_i (\boldsymbol{s}_i - \bar{\boldsymbol{s}})(\boldsymbol{s}_i - \bar{\boldsymbol{s}})^T \boldsymbol{z} \\
&= \sum_{i=1}^m w_i \boldsymbol{z}^T (\boldsymbol{s}_i - \bar{\boldsymbol{s}})(\boldsymbol{s}_i - \bar{\boldsymbol{s}})^T \boldsymbol{z} \\
&= \sum_{i=1}^m w_i \big\{ \boldsymbol{z}^T (\boldsymbol{s}_i - \bar{\boldsymbol{s}}) \big\}^2 \geq 0,
\end{aligned}
$$

for all $\boldsymbol{z} \in \mathbb{R}^d$ such that $||\boldsymbol{z}|| > 0$. In addition, define $\boldsymbol{q}_i = \boldsymbol{s}_i - \bar{\boldsymbol{s}}$ and assume that

$$r = \mathrm{rank}\,[\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m] = d. \qquad (14)$$

Then $\hat{K}''(\boldsymbol{\lambda})$ is positive definite and $\hat{K}(\boldsymbol{\lambda})$ is strictly convex. In fact, suppose that there exists a non-zero vector $\boldsymbol{z}$ such that $\boldsymbol{z}^T \hat{K}''(\boldsymbol{\lambda}) \boldsymbol{z} = 0$, which implies $\boldsymbol{z}^T \boldsymbol{q}_i = 0$ for $i = 1, \ldots, m$. Given that $\boldsymbol{z}$ can be expressed as a linear combination of $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_m$, this would imply that

$$\boldsymbol{z}^T \boldsymbol{z} = (b_1 \boldsymbol{q}_1 + \cdots + b_m \boldsymbol{q}_m)^T \boldsymbol{z} = 0,$$

which contradicts the fact that $\boldsymbol{z}$ is a non-zero vector. Now, define

$$
\begin{aligned}
J \subset \big\{1, \ldots, m\big\} \text{ such that } \quad &\boldsymbol{\lambda}^T \boldsymbol{s}_i = \alpha > 0 \text{ for all } i \in J, \\
&\boldsymbol{\lambda}^T \boldsymbol{s}_i < \alpha \text{ for all } i \notin J,
\end{aligned}
$$

examination of (13) shows that for all i, j such that $j \in J, i \notin J$ we have

$$\lim_{c \to \infty} w_i = \frac{\lim_{c \to \infty} e^{c(\boldsymbol{\lambda}^T \boldsymbol{s}_i - \boldsymbol{\lambda}^T \boldsymbol{s}_j)}}{\lim_{c \to \infty} \sum_{k=1}^m e^{c(\boldsymbol{\lambda}^T \boldsymbol{s}_k - \boldsymbol{\lambda}^T \boldsymbol{s}_j)}} = \frac{0}{\mathrm{Card}(J)} = 0,$$

while for all i, j such that $i, j \in J$ we have

$$\lim_{c \to \infty} w_i = \frac{\lim_{c \to \infty} e^{c(\boldsymbol{\lambda}^T \boldsymbol{s}_i - \boldsymbol{\lambda}^T \boldsymbol{s}_j)}}{\lim_{c \to \infty} \sum_{k=1}^m e^{c(\boldsymbol{\lambda}^T \boldsymbol{s}_k - \boldsymbol{\lambda}^T \boldsymbol{s}_j)}} = \frac{1}{\text{Card}(J)}.$$

Hence

$$\lim_{c \to \infty} \bar{\boldsymbol{s}} = \lim_{c \to \infty} \hat{K}'(c\boldsymbol{\lambda}) = \lim_{c \to \infty} \sum_{i=1}^m w_i \boldsymbol{s}_i = \frac{1}{\text{Card}(J)} \sum_{i \in J}^m \boldsymbol{s}_i,$$

and

$$\lim_{c \to \infty} \boldsymbol{\lambda}^T \boldsymbol{q}_i = \lim_{c \to \infty} \boldsymbol{\lambda}^T (\boldsymbol{s}_i - \bar{\boldsymbol{s}}) = \boldsymbol{\lambda}^T \left\{ \boldsymbol{s}_i - \frac{1}{\text{Card}(J)} \sum_{i \in J} \boldsymbol{s}_i \right\} = \alpha - \alpha = 0,$$

for all $i \in J$. Finally, we choose $\boldsymbol{z} = \boldsymbol{\lambda}$ and obtain

$$\lim_{c \to \infty} \boldsymbol{\lambda}^T \hat{K}''(c\boldsymbol{\lambda}) \boldsymbol{\lambda} = \sum_{i=1}^m \lim_{c \to \infty} w_i \lim_{c \to \infty} \left(\boldsymbol{\lambda}^T \boldsymbol{q}_i \right)^2 = \frac{1}{\text{Card}(J)} \sum_{i \in J} \lim_{c \to \infty} \left(\boldsymbol{\lambda}^T \boldsymbol{q}_i \right)^2 = 0,$$

which implies that $\hat{K}(\boldsymbol{\lambda})$ is not strongly convex.

A.2. Mean squared errors of the CGF estimators

Firstly notice that, irrespective of the distribution of $\boldsymbol{S}$, $\hat{M}(\boldsymbol{\lambda})$ is unbiased. If $\boldsymbol{S}$ is normally distributed, $e^{\boldsymbol{\lambda}^T \boldsymbol{S}}$ follows a log-normal distribution and

$$M(\boldsymbol{\lambda}) = e^{\boldsymbol{\mu} + \frac{1}{2} \boldsymbol{\lambda}^T \boldsymbol{\Sigma} \boldsymbol{\lambda}}, \quad \text{var}\{\hat{M}(\boldsymbol{\lambda})\} = \frac{1}{m} \left(e^{\boldsymbol{\lambda}^T \boldsymbol{\Sigma} \boldsymbol{\lambda}} - 1 \right) e^{2\boldsymbol{\mu} + \boldsymbol{\lambda}^T \boldsymbol{\Sigma} \boldsymbol{\lambda}},$$

with the saddlepoint equation (3) being solved by

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{\Sigma}^{-1} (\boldsymbol{s} - \boldsymbol{\mu}). \tag{15}$$

In order to approximate the MSE of (4) as a function of $\boldsymbol{\lambda}$, we firstly approximate its expected value by Taylor expansion around $M(\boldsymbol{\lambda})$

$$
\begin{aligned}
\mathbb{E}\{\hat{K}(\boldsymbol{\lambda})\} &= \mathbb{E}\left[\log M(\boldsymbol{\lambda}) + \frac{1}{M(\boldsymbol{\lambda})} \{\hat{M}(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})\} \right. \\
&\quad \left. - \frac{1}{2M(\boldsymbol{\lambda})^2} \{\hat{M}(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})\}^2 + \cdots \right] \\
&= \log M(\boldsymbol{\lambda}) - \frac{1}{2M(\boldsymbol{\lambda})^2} \text{var}\{\hat{M}(\boldsymbol{\lambda})\} + O(m^{-2}).
\end{aligned}
$$

Similarly we have that

$$
\begin{aligned}
\mathbb{E}\{\hat{K}(\boldsymbol{\lambda})^2\} &= \mathbb{E}\left[\{\log M(\boldsymbol{\lambda})\}^2 + \frac{2\log\{M(\boldsymbol{\lambda})\}}{M(\boldsymbol{\lambda})} \{\hat{M}(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})\} \right. \\
&\quad \left. + \left\{ \frac{1}{M(\boldsymbol{\lambda})^2} - \frac{\log M(\boldsymbol{\lambda})}{M(\boldsymbol{\lambda})^2} \right\} \{\hat{M}(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})\}^2 + \cdots \right] \\
&= \{\log M(\boldsymbol{\lambda})\}^2 + \left\{ \frac{1}{M(\boldsymbol{\lambda})^2} - \frac{\log M(\boldsymbol{\lambda})}{M(\boldsymbol{\lambda})^2} \right\} \text{var}\{\hat{M}(\boldsymbol{\lambda})\} + O(m^{-2}),
\end{aligned}
$$

hence

$$
\begin{aligned}
\text{var}\{\hat{K}(\boldsymbol{\lambda})\} &= \mathbb{E}\{\hat{K}(\boldsymbol{\lambda})^2\} - \mathbb{E}\{\hat{K}(\boldsymbol{\lambda})\}^2 \\
&= \frac{1}{M(\boldsymbol{\lambda})^2}\text{var}\{\hat{M}(\boldsymbol{\lambda})\} - \frac{1}{4M(\boldsymbol{\lambda})^4}\left[\text{var}\{\hat{M}(\boldsymbol{\lambda})\}\right]^2 + O(m^{-2}).
\end{aligned}
$$

Finally

$$
\begin{aligned}
\text{MSE}\{\hat{K}(\boldsymbol{\lambda})\} &= \text{Bias}\{\hat{K}(\boldsymbol{\lambda})\}^2 + \text{var}\{\hat{K}(\boldsymbol{\lambda})\} \\
&= \frac{1}{M(\boldsymbol{\lambda})^2}\text{var}\{\hat{M}(\boldsymbol{\lambda})\} + O(m^{-2}) \\
&= \frac{1}{m}(e^{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda}} - 1) + O(m^{-2}) \\
&= \frac{1}{m}\{e^{(\boldsymbol{s}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{s}-\boldsymbol{\mu})} - 1\} + O(m^{-2}),
\end{aligned}
\tag{16}
$$

where the last equality holds due to (15). The $O(m^{-2})$ term in (16) derives from

$$
\begin{aligned}
\mathbb{E}\left[\left\{\hat{M}(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})\right\}^3\right] &= \mathbb{E}\left[\left\{\frac{1}{m}\sum_{i=1}^m e^{\boldsymbol{\lambda}^T\boldsymbol{S}_i} - \mathbb{E}(e^{\boldsymbol{\lambda}^T\boldsymbol{S}})\right\}^3\right] \\
&= \frac{1}{m^3}\sum_{i=1}^m \mathbb{E}\left[\left\{e^{\boldsymbol{\lambda}^T\boldsymbol{S}_i} - \mathbb{E}(e^{\boldsymbol{\lambda}^T\boldsymbol{S}})\right\}^3\right] \\
&= \frac{1}{m^2}\mu_3(e^{\boldsymbol{\lambda}^T\boldsymbol{S}}),
\end{aligned}
$$

where $\mu_3(X)$ is the third central moment of a random variable X and the second equality is justified by independence.

Estimator (10) is unbiased and consistent, if $\boldsymbol{S}$ is normally distributed, hence

$$
\text{MSE}\{\hat{G}_m(\boldsymbol{\lambda})\} = \text{var}\{\hat{G}_m(\boldsymbol{\lambda})\} = \boldsymbol{\lambda}^T\text{var}(\hat{\boldsymbol{\mu}})\boldsymbol{\lambda} + \frac{1}{4}\text{var}\left(\boldsymbol{\lambda}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{\lambda}\right),
$$

due to the independence between $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ for normally distributed random variables (Basu's theorem). In addition, as m goes to infinity we have, from Rencher and Christensen (2012), that

$$
(m-1)\hat{\boldsymbol{\Sigma}} = \sum_{i=1}^m (\boldsymbol{S}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{S}_i - \hat{\boldsymbol{\mu}})^T \to \boldsymbol{W}, \quad \text{where} \quad \boldsymbol{W} \sim \text{Wishart}(\boldsymbol{\Sigma}, m-1),
$$

and from Rao (2009)

$$
\boldsymbol{\lambda}^T\boldsymbol{W}\boldsymbol{\lambda} \sim \tau^2 Q, \quad \text{where} \quad \tau^2 = \boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda} \quad \text{and} \quad Q \sim \chi^2_{m-1},
$$

hence, by using (15), we obtain

$$
\begin{aligned}
m\text{MSE}\{\hat{G}_m(\boldsymbol{\lambda})\} &\to \hat{\boldsymbol{\lambda}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\lambda}} + \frac{m}{2(m-1)}(\hat{\boldsymbol{\lambda}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\lambda}})^2 \\
&\to \hat{\boldsymbol{\lambda}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\lambda}}\left(1 + \frac{1}{2}\hat{\boldsymbol{\lambda}}^T\boldsymbol{\Sigma}\hat{\boldsymbol{\lambda}}\right) \\
&= (\boldsymbol{s}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{s}-\boldsymbol{\mu})\left\{1 + \frac{1}{2}(\boldsymbol{s}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{s}-\boldsymbol{\mu})\right\}.
\end{aligned}
$$

A.3. Proof of Theorem 1

The Gaussian synthetic log-likelihood is proportional to

$$n^{-\delta} \log \hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta}) \propto -(\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n)^T \left(n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n\right)^{-1} (\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n) - n^{-\delta} \log \det(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n). \quad (17)$$

Assumption 1 implies that

$$\begin{aligned} n^{-\delta} \log \det(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n n^{-\delta} n^{\delta}) &= n^{-\delta}\big\{ \log \det(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n n^{\delta}) - d\delta \log n \big\} \\ &= O_p(n^{-\delta}) + O(n^{-\delta} \log n), \end{aligned}$$

so, as n and $m \to \infty$, the r.h.s. of (17) converges in probability to

$$f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) = -(\boldsymbol{\mu}_{\boldsymbol{\theta}_0} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\mu}_{\boldsymbol{\theta}_0} - \boldsymbol{\mu}_{\boldsymbol{\theta}}),$$

for any $\boldsymbol{\theta}$. Here $\boldsymbol{\mu}_{\boldsymbol{\theta}_0}$ is the asymptotic mean vector at true parameters $\boldsymbol{\theta}_0$. Then

$$f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) = -(\boldsymbol{\mu}_{\boldsymbol{\theta}_0} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T \boldsymbol{U}_{\boldsymbol{\theta}} \boldsymbol{\Psi}_{\boldsymbol{\theta}}^{-1} \boldsymbol{U}_{\boldsymbol{\theta}}^T (\boldsymbol{\mu}_{\boldsymbol{\theta}_0} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) = -\boldsymbol{z}_{\boldsymbol{\theta}}^T \boldsymbol{\Psi}_{\boldsymbol{\theta}}^{-1} \boldsymbol{z}_{\boldsymbol{\theta}},$$

where $\boldsymbol{U}_{\boldsymbol{\theta}} \boldsymbol{\Psi}_{\boldsymbol{\theta}} \boldsymbol{U}_{\boldsymbol{\theta}}^T$ is the eigen-decomposition of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, and we defined $\boldsymbol{z}_{\boldsymbol{\theta}} = \boldsymbol{U}_{\boldsymbol{\theta}}^T (\boldsymbol{\mu}_{\boldsymbol{\theta}_0} - \boldsymbol{\mu}_{\boldsymbol{\theta}})$. Now, if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, then Assumption 3 assures that $||\boldsymbol{z}_{\boldsymbol{\theta}}||_2 = ||\boldsymbol{\mu}_{\boldsymbol{\theta}_0} - \boldsymbol{\mu}_{\boldsymbol{\theta}}||_2 > 0$ which, together with Assumption 2, guarantees that

$$f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) = -\sum_{i=1}^{d} \frac{1}{(\boldsymbol{\Psi}_{\boldsymbol{\theta}})_{ii}}(\boldsymbol{z}_{\boldsymbol{\theta}})_i^2 \leq -\frac{1}{{}_*\psi}||\boldsymbol{z}_{\boldsymbol{\theta}}||_2^2 < 0.$$

Given that $f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}_0) = 0$, this function is maximized at $\boldsymbol{\theta}_0$, which implies identifiability under a Gaussian density estimator.

A.4. Proof of Theorem 2

Given Assumption 5 and the fact that all the functions involved in $n^{-\delta} \log \hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta})$ are continuously differentiable, this function is continuously differentiable itself, due to the chain rule. We then have to show that its derivative is dominated by an $O_p(1)$ random sequence. Consider the partial derivative of the log-determinant w.r.t. the k-th parameter

$$\left| \frac{\partial \log \det(n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n)}{\partial \theta_k} \right| = \left| \mathrm{Tr}\left[(n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n)^{-1} \frac{\partial n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n}{\partial \theta_k} \right] \right|$$

$$\leq \mathrm{Tr}\left[\left|\hat{\boldsymbol{U}}_{\boldsymbol{\theta}}^n\right| \left|(\hat{\boldsymbol{\Psi}}_{\boldsymbol{\theta}}^n)^{-1}\right| \left|(\hat{\boldsymbol{U}}_{\boldsymbol{\theta}}^n)^T\right| \left|\frac{\partial n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n}{\partial \theta_k}\right| \right],$$

where $\hat{\boldsymbol{U}}_{\boldsymbol{\theta}}^n \hat{\boldsymbol{\Psi}}_{\boldsymbol{\theta}}^n (\hat{\boldsymbol{U}}_{\boldsymbol{\theta}}^n)^T$ is the eigen-decomposition of $n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n$. Then

$$\left| \frac{\partial \log \det(n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n)}{\partial \theta_k} \right| \leq \frac{1}{{}_*\hat{\psi}_{\boldsymbol{\theta}}^n} \mathrm{Tr}\left[\left|\hat{\boldsymbol{U}}_{\boldsymbol{\theta}}^n\right| \left|(\hat{\boldsymbol{U}}_{\boldsymbol{\theta}}^n)^T\right| \left|\frac{\partial n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n}{\partial \theta_k}\right| \right] \leq \frac{d}{{}_*\hat{\psi}_{\boldsymbol{\theta}}^n} \sum_{i=1}^{d} \sum_{j=1}^{d} \left|\frac{\partial n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n}{\partial \theta_k}\right|_{ij}.$$

$$(18)$$

Now, consider the derivative of the inverse (scaled) covariance matrix

$$\left|\left|\frac{\partial\left(n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{n}\right)^{-1}}{\partial\theta_k}\right|\right|_2 = \left|\left|\left(n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{n}\right)^{-1}\frac{\partial n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{n}}{\partial\theta_k}\left(n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{n}\right)^{-1}\right|\right|_2$$
$$\leq \left|\left|\frac{\partial n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{n}}{\partial\theta_k}\right|\right|_2\left|\left|\left(n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{n}\right)^{-1}\right|\right|_2^2,$$

but

$$\left|\left|\left(n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{n}\right)^{-1}\right|\right|_2 = \left|\left|\hat{\boldsymbol{U}}_{\boldsymbol{\theta}}^{n}\right|\right|_2\left|\left|\left(\hat{\boldsymbol{\Psi}}_{\boldsymbol{\theta}}^{n}\right)^{-1}\right|\right|_2\left|\left|\left(\hat{\boldsymbol{U}}_{\boldsymbol{\theta}}^{n}\right)^{T}\right|\right|_2 \leq \frac{1}{{}_*\hat{\psi}_{\boldsymbol{\theta}}^{n}},$$

hence

$$\left|\left|\frac{\partial\left(n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{n}\right)^{-1}}{\partial\theta_k}\right|\right|_2 \leq \left({}_*\hat{\psi}_{\boldsymbol{\theta}}^{n}\right)^{-2}\left|\left|\frac{\partial n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{n}}{\partial\theta_k}\right|\right|_2. \tag{19}$$

Under Assumptions 5 and 6, the r.h.s. of both (18) and (19) are dominated. So if we consider

$$\left|\frac{\partial n^{-\delta}\log\hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta})}{\partial\theta_k}\right| \leq \left|\left|\frac{\partial\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{n}}{\partial\theta_k}\right|\right|_2||n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{n}||_2||\boldsymbol{s}_0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{n}||_2$$
$$+ \frac{1}{2}\left|\left|\frac{\partial\left(n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{n}\right)^{-1}}{\partial\theta_k}\right|\right|_2||\boldsymbol{s}_0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{n}||_2^2 + \frac{1}{2}\left|\frac{\partial\log\det(n^{\delta}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{n})}{\partial\theta_k}\right|, \tag{20}$$

it is clear that the r.h.s. of (20) is dominated as well, provided that $||\boldsymbol{s}_0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{n}||_2$ is. Now

$$||\boldsymbol{s}_0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{n}||_2^2 = ||\boldsymbol{z}_{\boldsymbol{\theta}_0}^{n} + \boldsymbol{\mu}_{\boldsymbol{\theta}_0}^{n} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_0}^{n} + \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_0}^{n} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{n}||_2^2$$
$$\leq ||\boldsymbol{z}_{\boldsymbol{\theta}_0}^{n}||_2^2 + ||\boldsymbol{\mu}_{\boldsymbol{\theta}_0}^{n} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_0}^{n}||_2^2 + ||\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_0}^{n} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{n}||_2^2,$$

where $\boldsymbol{z}_{\boldsymbol{\theta}_0}^{n} = \boldsymbol{s}_0 - \boldsymbol{\mu}_{\boldsymbol{\theta}_0}^{n}$. But $||\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_0}^{n} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{n}||_2^2$ is dominated, because the derivatives of $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^{n}$ are dominated by Assumption 5 and the parameter space is compact by Assumption 4. In addition, for any $\epsilon > 0$, by Markov's inequality

$$\text{Prob}\left(||\boldsymbol{z}_{\boldsymbol{\theta}_0}^{n}||_2^2 > \epsilon\right) \leq \frac{\mathbb{E}\left(||\boldsymbol{z}_{\boldsymbol{\theta}_0}^{n}||_2^2\right)}{\epsilon} = n^{-\delta}\frac{\text{Tr}\left(n^{\delta}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{n}\right)}{\epsilon} \leq \frac{n^{-\delta}\,d\,{}^{*}\psi_{\boldsymbol{\theta}_0}^{n}}{\epsilon}, \tag{21}$$

where ${}^{*}\psi_{\boldsymbol{\theta}_0}^{n}$ is the largest eigenvalue of $n^{\delta}\boldsymbol{\Sigma}_{\boldsymbol{\theta}_0}^{n}$. The r.h.s. of (21) is $O(n^{-\delta})$ by Assumptions 1 and 6, hence $||\boldsymbol{z}_{\boldsymbol{\theta}_0}^{n}||_2^2$ is $o_p(1)$. An identical argument shows that $||\boldsymbol{\mu}_{\boldsymbol{\theta}_0}^{n} - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}_0}^{n}||_2^2$ is $o_p(1)$. This proves that the derivatives of $n^{-\delta}\log\hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta})$ are dominated by an $O_p(1)$ sequence.

As explained in Section 4, this implies the uniform convergence of $n^{-\delta}\log\hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta})$ to $f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$, provided that $f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$ is equicontinuous. It is easy to show to that, if the derivatives of $f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$ are bounded, then equicontinuity follows. But, under Assumptions 2 and 7, it is possible to bound the derivatives of $f_{\boldsymbol{\theta}_0}(\boldsymbol{\theta})$, as we have just done for $n^{-\delta}\log\hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta})$. This assures uniform convergence which, together with identifiability, guarantees weak consistency (Van der Vaart, 2000).

A.5. Proof of Theorem 3

Taylor expanding the un-normalized EES-based synthetic log-likelihood leads to

$$\log \hat{p}_S(\boldsymbol{s}^0|\boldsymbol{\theta}) = \log \hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta}) + O\big\{e^{-\hat{\gamma}_{\boldsymbol{\theta}}^n \, (\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n)^T \left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n\right)^{-1} (\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n)}\big\}, \tag{22}$$

and, by multiplying both sides by $n^{-\delta}$, we obtain

$$n^{-\delta} \log \hat{p}_S(\boldsymbol{s}^0|\boldsymbol{\theta}) = n^{-\delta} \log \hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta}) + O\big\{n^{-\delta} e^{-\hat{\gamma}_{\boldsymbol{\theta}}^n n^\delta \, (\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n)^T \left(n^\delta \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^n\right)^{-1} (\boldsymbol{s}^0 - \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}^n)}\big\},$$

and, as m and $n \to \infty$, Assumptions 1, 2 and 9 imply

$$n^{-\delta} \log \hat{p}_S(\boldsymbol{s}^0|\boldsymbol{\theta}) \to n^{-\delta} \log \hat{p}_G(\boldsymbol{s}^0|\boldsymbol{\theta}), \tag{23}$$

in probability, for any $\boldsymbol{\theta}$. Identifiability then follows from Theorem 1.

Appendix B: Further theoretical results

B.1. Asymptotics of the multivariate empirical saddlepoint approximation

Here we follow Feuerverger (1989) but develop the results in a multivariate setting, and with some changes in notation. For $\boldsymbol{\lambda} \in I$, $\hat{M}_m(\boldsymbol{\lambda})$ converges to $M(\boldsymbol{\lambda})$ almost surely. This convergence is uniform and extends to $\hat{K}_m(\boldsymbol{\lambda})$:

$$\sup_{\boldsymbol{\lambda} \in I} |\hat{M}_m(\boldsymbol{\lambda}) - M(\boldsymbol{\lambda})| \to 0, \tag{24}$$

$$\sup_{\boldsymbol{\lambda} \in I} |\hat{K}_m(\boldsymbol{\lambda}) - K(\boldsymbol{\lambda})| \to 0. \tag{25}$$

The proof is as follows. Due to the Strong Law of Large Numbers $\hat{M}_m(\boldsymbol{\lambda})$ converges to $M(\boldsymbol{\lambda})$ almost surely, for all $\boldsymbol{\lambda}$ in any countable collection $\{\boldsymbol{\lambda}_i\}$. In addition $\hat{M}_m(\boldsymbol{\lambda})$ and $M(\boldsymbol{\lambda})$ are both convex functions and, for such functions, convergence on dense subsets implies uniform convergence on compact subsets (Roberts and Varberg, 1973). This proves (24), while (25) follows by continuity of the logarithm.

For $\boldsymbol{\lambda}$ in the interior of I, these results extend to derivatives of both $\hat{M}_m(\boldsymbol{\lambda})$ and $\hat{K}_m(\boldsymbol{\lambda})$:

$$\sup_{\boldsymbol{\lambda} \in \mathrm{int}(I)} |D^i \hat{M}_m(\boldsymbol{\lambda}) - D^i M(\boldsymbol{\lambda})| \to 0, \tag{26}$$

$$\sup_{\boldsymbol{\lambda} \in \mathrm{int}(I)} |D^i \hat{K}_m(\boldsymbol{\lambda}) - D^i K(\boldsymbol{\lambda})| \to 0, \tag{27}$$

where $i = \{i_1, \ldots, i_d\}$ and:

$$D^i M(\boldsymbol{\lambda}) = \frac{\partial^k M(\boldsymbol{\lambda})}{\partial \lambda_1^{i_1} \cdots \partial \lambda_d^{i_d}}, \quad \text{with} \quad \sum_{z=1}^K i_z = k \in N.$$

The proof is as follows. $D^i M(\boldsymbol{\lambda})$ is finite only for $\boldsymbol{\lambda} \in \mathrm{int}(I)$. If all the elements of i are even, then $D^i \hat{M}_m(\boldsymbol{\lambda})$ and $D^i M(\boldsymbol{\lambda})$ are convex and (26) follows as before. Otherwise, indicate with $\boldsymbol{\lambda}^o$ the elements of $\boldsymbol{\lambda}$ for which the corresponding element of i is odd. If there is an even number of components of $\boldsymbol{\lambda}^o$ which are negative, $D^i M(\boldsymbol{\lambda})$ is still convex, otherwise $-D^i M(\boldsymbol{\lambda})$ is. Applying the uniform convergence argument for convex functions to the two sub-cases proves (26). In addition, $D^i K(\boldsymbol{\lambda})$ has the form $P(\boldsymbol{\lambda})/M(\boldsymbol{\lambda})^{2^k}$ with $P(\boldsymbol{\lambda})$ being a polynomial function of $D^l K(\boldsymbol{\lambda})$, where l belongs to the set of all d-dimensional vector such that:

$$l_j \in N, \quad \sum_{j=1}^{d} l_j \leq k \quad \text{for} \quad j = 1, \ldots, d.$$

Given that an analogous argument holds for $D^i \hat{K}_m(\boldsymbol{\lambda})$, (27) is proved by continuity.

After noticing that $\hat{M}_m(\boldsymbol{\lambda})$ and its derivatives are unbiased estimators of $M(\boldsymbol{\lambda})$ and its corresponding derivatives, it is straightforward to show that:

$$m \, \mathrm{Cov}\left\{D^i \hat{M}_m(\boldsymbol{\lambda}_1), D^j \hat{M}_m(\boldsymbol{\lambda}_2)\right\} = D^{i+j} M(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2) - D^i M(\boldsymbol{\lambda}_1) D^j M(\boldsymbol{\lambda}_2),$$

for $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$ such that $\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2 \in I$. This entails that, if we define $I/2$ to be the subset of I such that $\boldsymbol{\lambda} \in I/2$ if $2\boldsymbol{\lambda} \in I$, than $\hat{M}_m(\boldsymbol{\lambda})$ is a $\sqrt{m}$-consistent estimator of $M(\boldsymbol{\lambda})$, for $\boldsymbol{\lambda} \in I/2$. An analogous, but asymptotic, result for $\hat{K}_m(\boldsymbol{\lambda})$ is the following:

$$m \, \mathrm{Cov}\left\{D^i \hat{K}_m(\boldsymbol{\lambda}_1), D^j \hat{K}_m(\boldsymbol{\lambda}_2)\right\} \to D^{i+j}\left\{\frac{M(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)}{M(\boldsymbol{\lambda}_1) M(\boldsymbol{\lambda}_2)} - 1\right\},$$

where $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are further restricted to the interior of $I/2$ if any of the elements of i or j is greater than zero. Finally, after noticing that on $I/2$:

$$\hat{\boldsymbol{\lambda}} = \hat{K'}^{-1}(\boldsymbol{x}) = \boldsymbol{\lambda} + O(m^{-\frac{1}{2}}),$$

we have that:

$$
\begin{aligned}
\frac{\hat{p}_m(\boldsymbol{s})}{\hat{p}(\boldsymbol{s})} &= \frac{\det\{K''(\boldsymbol{\lambda})\}}{\det\{\hat{K}''_m(\hat{\boldsymbol{\lambda}})\}} \exp\left[\left\{\hat{K}_m(\hat{\boldsymbol{\lambda}}) - \hat{\boldsymbol{\lambda}}^T \hat{K}'_m(\hat{\boldsymbol{\lambda}})\right\} - \left\{K(\boldsymbol{\lambda}) - \boldsymbol{\lambda}^T K'(\boldsymbol{\lambda})\right\}\right] \\
&= \frac{\det\{K''(\boldsymbol{\lambda})\}}{\det\{K''(\boldsymbol{\lambda})\} + O(m^{-\frac{1}{2}})} \exp\left\{O(m^{-1/2})\right\} \\
&= 1 + O(m^{-\frac{1}{2}}),
\end{aligned}
$$

by Taylor expansions, which are justified by the differentiability of all the functions involved. See Feuerverger (1989) for more details.

B.2. *Optimality of the cross-validated extended empirical saddlepoint*

Let $p(\boldsymbol{s}|\boldsymbol{\theta})$ be the true density of the statistics and $\hat{p}_S(\boldsymbol{s}|\boldsymbol{\theta}, \gamma)$ be the EES density. Assume that we have a training set of size m, a test set of size n_T and that we

have used l simulations to normalize the density estimator. In this section we prove that, as m, n_T and $l \to \infty$, Algorithm 2 consistently selects the value of γ which minimizes the Kullback-Leibler divergence between $\hat{p}_S(\boldsymbol{s}|\boldsymbol{\theta}, \gamma)$ and $p(\boldsymbol{s}|\boldsymbol{\theta})$. When two folds are used, cross-validation (Algorithm 2) selects γ as follows

$$\hat{\gamma} = \underset{\gamma}{\operatorname{argmin}} \left\{ -\frac{1}{n_T} \sum_{i=1}^{n_T} \log \hat{p}_S(\boldsymbol{s}_i|\boldsymbol{\theta}, \gamma) \right\} \quad \text{with} \quad \boldsymbol{s}_i \sim p(\boldsymbol{s}|\boldsymbol{\theta}),$$

but the Weak Law of Large Numbers implies that

$$
\begin{aligned}
\underset{m,l,n_T \to \infty}{\operatorname{plim}} -\tfrac{1}{n_T} \sum_{i=1}^{n_T} \log \hat{p}_S(\boldsymbol{s}_i|\boldsymbol{\theta}, \gamma) &= -\int \log p_S(\boldsymbol{s}|\boldsymbol{\theta}, \gamma) p(\boldsymbol{s}|\boldsymbol{\theta}) ds \\
&\propto \int \log \tfrac{p(\boldsymbol{s}|\boldsymbol{\theta})}{p_S(\boldsymbol{s}|\boldsymbol{\theta}, \gamma)} p(\boldsymbol{s}|\boldsymbol{\theta}) ds \\
&= \operatorname{KL}\{p_S(\boldsymbol{s}|\boldsymbol{\theta}, \gamma), p(\boldsymbol{s}|\boldsymbol{\theta})\}.
\end{aligned}
$$

Hence $p_S(\boldsymbol{s}|\boldsymbol{\theta}, \hat{\gamma})$ is the member of the $p_S(\boldsymbol{s}|\boldsymbol{\theta}, \gamma)$ family with minimal Kullback-Leibler distance from $p(\boldsymbol{s}|\boldsymbol{\theta})$. This result can easily be extended to k-fold cross-validation ($k > 2$).

Appendix C: Details on practical implementation and examples

C.1. Saddlepoint version of Algorithm 1

In this section we illustrate how a pointwise synthetic likelihood estimate can be obtained using the new density estimator, rather than a Gaussian density.

C.2. The iterated filtering optimizer

Here we describe the basic iteration performed by the IF procedure. Suppose that $\hat{\boldsymbol{\theta}}_k$ is the estimate of the unknown parameters at the k-th step of the optimization routine. This estimate is updated as follows:

1. Simulate N parameter vectors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ from a user-defined density $p(\boldsymbol{\theta}_{k+1}|\hat{\boldsymbol{\theta}}_k)$ such that

$$\mathbb{E}(\boldsymbol{\theta}_{k+1}|\hat{\boldsymbol{\theta}}_k) = \hat{\boldsymbol{\theta}}_k, \quad \operatorname{var}(\boldsymbol{\theta}_{k+1}|\hat{\boldsymbol{\theta}}_k) = \sigma_k^2 \boldsymbol{\Sigma} \quad \text{and} \quad \mathbb{E}(||\boldsymbol{\theta}_{k+1} - \hat{\boldsymbol{\theta}}_k||^{3/2}) = o(\sigma_k^2), \tag{28}$$

   where σ_k^2 is a cooling schedule and $\boldsymbol{\Sigma}$ is a covariance matrix.
2. For each $\boldsymbol{\theta}_i$, obtain an estimate $\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}_i)$ of the synthetic likelihood, using either the multivariate normal density or the normalized EES.
3. Update the estimate

$$\hat{\boldsymbol{\theta}}_{k+1} = \frac{\sum_{i=1}^{N} \boldsymbol{\theta}_i \hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}_i)}{\sum_{i=1}^{N} \hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}_i)}.$$

Algorithm 3 Estimating $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ using EES

1: Simulate datasets $\boldsymbol{Y}_i,\ldots,\boldsymbol{Y}_m$ from the model $p(\boldsymbol{Y}|\boldsymbol{\theta})$.
2: Transform each dataset $\boldsymbol{Y}_i$ to a vector of summary statistics $\boldsymbol{S}_i = S(\boldsymbol{Y}_i)$.
3: Calculate sample mean $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ of the simulated statistics.
4: Estimate the synthetic likelihood

$$\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}) = \hat{p}_m(\boldsymbol{s}^0,\gamma) = \frac{1}{(2\pi)^{\frac{d}{2}}\det\{\hat{K}_m''(\hat{\boldsymbol{\lambda}}_m,\gamma,\boldsymbol{s}^0)\}^{\frac{1}{2}}}e^{\hat{K}_m(\hat{\boldsymbol{\lambda}}_m,\gamma,\boldsymbol{s}^0)-\hat{\boldsymbol{\lambda}}_m^T\boldsymbol{s}^0},$$

where $\hat{\boldsymbol{\lambda}}_m$ is the solution of the empirical saddlepoint equation

$$\hat{K}_m'(\hat{\boldsymbol{\lambda}}_m,\gamma,\boldsymbol{s}^0) = \boldsymbol{s}^0,$$

while $\hat{K}_m(\boldsymbol{\lambda},\gamma,\boldsymbol{s})$ is given by equation (8) in the main text.
5: Optionally, normalize $\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ by importance sampling

$$\bar{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta}) = \frac{\hat{p}_m(\boldsymbol{s}^0,\gamma)}{\hat{z}_m(\gamma)},$$

where

$$\hat{z}_m(\gamma) = \frac{1}{l}\sum_{i=1}^{l}\frac{\hat{p}_m(\boldsymbol{S}_i,\gamma)}{q(\boldsymbol{S}_i)},\quad \boldsymbol{S}_i \sim q(\boldsymbol{s}),\quad \text{for}\ \ i=1,\ldots,l.$$

A reasonably efficient importance density $q(\boldsymbol{s})$ is a Gaussian density with mean vector $\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}$ and covariance $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$.
6: Output: the unnormalized, $\hat{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$, or the normalized, $\bar{p}_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$, synthetic likelihood.

Hence $\hat{\boldsymbol{\theta}}_{k+1}$ is a convex combination of $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N$. For all the examples we used the following cooling schedule

$$\sigma_k^2 = \sigma_0^{2k},\quad \sigma_0^2 = 0.95.$$

In the shifted exponential example we performed 4 separate runs of the optimizer, using either the normal or the EES approximation, in both the 10 and 20-dimensional setting.

C.3. Shifted exponential details

In one dimension, the ABC likelihood is

$$p_\epsilon(s^0|\theta) = \int_0^\infty I(|s-s^0|<\epsilon)p(s|\theta)ds,$$

Without loss of generality, choose $s^0 = 0$. If $-\epsilon \le \theta \le \epsilon$ we have

$$p_\epsilon(s^0|\theta) = \int_\theta^\epsilon p(s|\theta)ds = \int_\theta^\epsilon \beta e^{-\beta(s-\theta)}ds = \int_0^{\epsilon-\theta}\beta e^{-\beta x}dx = 1 - e^{-\beta(\epsilon-\theta)}$$
$$= F(\epsilon-\theta),$$

where we used the change of variable $x = s - \theta$. Similarly, if $\theta < -\epsilon$, the likelihood is

$$p_\epsilon(s^0|\theta) = \int_{-\epsilon}^\epsilon p(s|\theta)ds = F(\epsilon-\theta) - F(-\epsilon-\theta) = e^{-\beta(-\epsilon-\theta)} - e^{-\beta(\epsilon-\theta)}.$$

Finally, $p_\epsilon(s^0|\theta) = 0$ for $\theta > \epsilon$. Under a uniform prior on $[\psi, 0]$, where $\psi < -\epsilon$, the MAP is

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p_\epsilon(s^0|\theta) = -\epsilon,$$

with $p_\epsilon(s^0|\hat{\theta}) = F(2\epsilon) = 1 - e^{-2\beta\epsilon}$. In d dimensions, the likelihood is

$$p_\epsilon(\boldsymbol{s}^0|\boldsymbol{\theta}) = \prod_{k=1}^{d} p_\epsilon(s_k^0|\theta_k),$$

due to the independence between the summary statistics. Hence, the likelihood at the MLE is $F(2\epsilon)^d$, which is also the maximal probability that a simulated statistics vector gets accepted.

In ABC the tolerance is often chosen so that a fraction $\alpha \in (0,1)$ of the statistics simulated from the prior falls within the tolerance. In one dimension and for fixed ψ, the overall probability of acceptance during this process is

$$
\begin{aligned}
p(|S| < \epsilon|\psi) &= \int_\psi^0 p(-\epsilon < S < \epsilon|\theta)\frac{1}{-\psi}d\theta \\
&= -\frac{1}{\psi}\left\{ \int_\psi^{-\epsilon} \left[e^{-\beta(-\epsilon-\theta)} - e^{-\beta(\epsilon-\theta)} \right]d\theta + \int_{-\epsilon}^0 (1 - e^{-\beta(\epsilon-\theta)})d\theta \right\} \\
&= -\frac{1}{\psi}\left\{ \frac{1}{\beta}\left[1 + e^{\beta\psi}\left(e^{-\beta\epsilon} - e^{\beta\epsilon} \right) - e^{-\beta\epsilon} \right] + \epsilon \right\}.
\end{aligned}
$$

Now, to select ψ so that we obtain an acceptance probability equal to ϕ, we need to solve

$$p(|S| < \epsilon|\psi) = \phi,$$

wrt ψ, numerically (e.g. using bisection). Due to the independence between the priors and between the statistics, in d dimensions the above probability becomes $p(|S| < \epsilon|\psi)^d$.

C.4. Unstable population model details

Under the Gaussian and EES version of SL, we maximize the synthetic likelihood using 100 iterations of Iterated Filtering, with $N = 24$ synthetic likelihood evaluations at each step. The optimizer is initialized at $r = 0.3$, $\kappa = 30$, $\alpha = 0.15$ and $\beta = 0.03$. Given that $m = 5 \times 10^3$, estimating the model parameters costs 12×10^6 simulations from the model. In ABC we use 10^6 simulation to calibrate the tolerance ϵ, followed by 12×10^6 MCMC samples. Of these, we store only a thinned sub-sample of 22×10^3 parameter vector. Before obtaining MAP estimates, we discard the first 4×10^3 of these as burn-in. Then we maximize the posterior using the mean shift algorithm, where the approximate posterior is estimated using a Gaussian kernel density estimator. Following the rule of thumb of Silverman (1986) the covariance matrix, $\mathbf{H}$, of the kernels is determined as follows

$$\mathbf{H}^{\frac{1}{2}} = \left(\frac{4}{d+2} \right)^{\frac{1}{d+4}} n^{-\frac{1}{d+4}} \hat{\boldsymbol{\Sigma}}^{1/2},$$

where $n = 18 \times 10^3$, while $\mathbf{H}^{\frac{1}{2}}$ and $\boldsymbol{\Sigma}^{1/2}$ are matrix square roots of $\mathbf{H}$ and of $\hat{\boldsymbol{\Sigma}}$, the estimated covariance of the posterior samples. Given that the kernel density estimate of the posterior might have multiple local modes, we obtain 500 different MAP estimates by initializing the mean shift algorithm at a random posterior sample. We used the estimate corresponding to the highest (estimate) posterior density as our final MAP estimate.

C.5. Formind settings

The summary statistic were obtained using the following constants

$$\alpha_{1,1} = \alpha_{1,3} = \alpha_{2,1} = \alpha_{2,3} = 1.5, \ \alpha_{1,2} = 2, \alpha_{2,2} = 2$$

while ψ_{jk} and σ_{jk} were estimates of mean and standard deviations of C_{jk}, obtained by simulating tree counts at the true parameters. The 24 datasets were simulated from Formind using the same parameter values as in Table 1 in the supplementary material of Hartig et al. (2014). The chosen tree classes correspond to diameters at breast height $d < 0.2m$, $0.2m \leq d < 0.6m$, $d \geq 0.6m$ for pioneer and $d < 0.5m$, $0.5m \leq d < 1.4m$, $d \geq 1.4m$ for late successional trees. To generate the datasets the model was run for 10^5 years, and the final statistics vector was selected. The $m = 10^4$ summary statistics simulated to estimate $p_{SL}(\boldsymbol{s}^0|\boldsymbol{\theta})$ have been generated by simulating the model for 5.1×10^4 years, where the first 10^3 years of simulation were discarded to avoid the transient, and by storing a vector of statistics every 5 years.

Starting from initial values $\mu_{pio} = 0.03$, $\mu_{suc} = 0.003$, $s_{pio} = 120$ and $s_{suc} = 40$, we ran the optimizations using $N = 24$ and 100 iterations. The estimates reported in Table 1 in the main text were obtained by using the averages of the last 10 iterations of each optimization run as point estimates. The whole experiment took around 10 days on a quad-core Intel i7 3.6 GHz processor.

C.6. A further example: correlated multivariate shifted exponential distribution

In the shifted exponential example included in the main text, the elements of the random vector $\boldsymbol{S}$ are independent. To show that EES can cope with correlated random variables we have introduced correlations, without altering the marginal densities, by using a copula model. In particular, we used a Gaussian d-dimensional copula, which has density

$$c(u_1, \ldots, u_d|\boldsymbol{R}) = \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left\{\frac{1}{2}\boldsymbol{q}^T(\boldsymbol{I}_d - \boldsymbol{R}^{-1})\boldsymbol{q}\right\},$$

where $\boldsymbol{R}$ is a $d \times d$ correlation matrix, $\boldsymbol{I}_d$ is the identity matrix, $\boldsymbol{q}$ is a d-dimensional vector with $q_i = \Phi^{-1}(u^i)$, where Φ is the cumulative distribution function of a standard normal. The random vector $\{u_1, \ldots, u_d\}$ has marginals
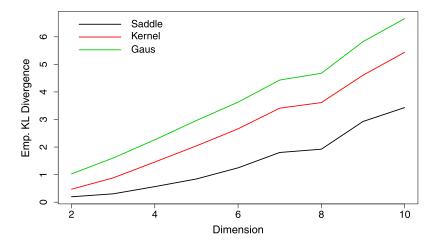
FIG 5. *Empirical Kullback-Leibler divergence between the three density estimators and the true density, as the number of dimensions increases.*

that are uniformly distributed on $[0, 1]$. For an introduction to copulas see Cherubini et al. (2004).

To simulate unstructured, dense correlation matrices $\boldsymbol{R}$, we have used the method proposed by Joe (2006). To set up the copula model and to simulate random variable, we have used the tools described by Yan et al. (2007).

We compare EES with a Gaussian and a kernel density estimator. In particular, we used $m = 10^3$ training samples and 5×10^3 test samples. The normalizing constant of the saddlepoint was estimated using $l = 10^3$ simulations and γ was estimated by cross-validation. For the kernel estimator we used a multivariate Gaussian kernel with covariance $\alpha \hat{\boldsymbol{\Sigma}}$, where $\hat{\boldsymbol{\Sigma}}$ is the empirical covariance matrix of the random vectors in training set and α is a scaling parameter, whose value was selected by cross-validation. Figure 5 shows how the estimated Kullback-Leibler divergence, between the true density and each density estimate, varies with the number of dimensions. The true density is very skewed in each dimension, hence the Gaussian estimator is highly biased. The kernel estimator does better than the Gaussian, even as the dimensionality increases. This is attributable to the fact that having a single bandwidth α is very helpful in this example, because all the marginal densities are identical. The new density estimator performs uniformly better than the alternatives.

As in the uncorrelated scenario (see the main text) we now estimate the shifts $\theta_1, \ldots, \theta_d$, using the Gaussian and the new density estimator. We have considered a 10 and a 20-dimensional scenario. In both cases γ has been selected by cross-validation. We have used $\beta = 0.2$ and $\theta_1 = \cdots = \theta_d = 0$. Given that the shape of the densities does not change with any of the θs we set $l = 0$, and we have not computed the normalizing constant. We have used $m = 10^4$ and $m = 5 \times 10^4$ simulated vectors, respectively. By using EES, the MSE was reduced from 21.9 to 4.8 in the 10-dimensional setting, and from 22.7 to 3.2 in

the 20-dimensional setting. P-values from t-test for differences in log-absolute errors were lower than 10^{-9} in both runs.

Acknowledgments

References

Andrieu, C., G. O. Roberts, et al. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics 37*(2), 697–725. MR2502648

Bartolucci, F. (2007). A penalized version of the empirical likelihood ratio for the population mean. *Statistics & Probability Letters 77*(1), 104–110. MR2339024

Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate Bayesian computation in population genetics. *Genetics 162*(4), 2025–2035.

Blum, M., M. Nunes, D. Prangle, and S. Sisson (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science 28*(2), 189–208. MR3112405

Blum, M. G. (2010). Approximate Bayesian computation: a nonparametric perspective. *Journal of the American Statistical Association 105*(491). MR2752613

Butler, R. W. (2007). *Saddlepoint approximations with applications*. Cambridge University Press. MR2357347

Cherubini, U., E. Luciano, and W. Vecchiato (2004). *Copula methods in finance*. John Wiley & Sons. MR2250804

Daniels, H. E. (1954). Saddlepoint approximations in statistics. *The Annals of Mathematical Statistics 25*(4), 631–650. MR0066602

Davison, A. C. and D. V. Hinkley (1988). Saddlepoint approximations in resampling methods. *Biometrika 75*(3), 417–431. MR0967581

Dislich, C., S. Günter, J. Homeier, B. Schröder, and A. Huth (2009). Simulating forest dynamics of a tropical montane forest in South Ecuador. *Erdkunde 63*(4), 347–364.

Doucet, A., S. Godsill, and C. Andrieu (2000). On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing 10*(3), 197–208.

Doucet, A., P. E. Jacob, and S. Rubenthaler (2013). Derivative-free estimation of the score vector and observed information matrix with application to state-space models. *arXiv preprint arXiv:1304.5768*.

Durbin, J. (1980). Approximations for densities of sufficient estimators. *Biometrika 67*(2), 311–333. MR0581728

Easton, G. S. and E. Ronchetti (1986). General saddlepoint approximations with applications to L-statistics. *Journal of the American Statistical Association 81*(394), 420–430. MR0845879

Everitt, R. G., A. M. Johansen, E. Rowing, and M. Evdemon-Hogan (2015). Bayesian model comparison with intractable likelihoods. *arXiv preprint arXiv:1504.00298*. MR3599680

Fasiolo, M., N. Pya, S. N. Wood, et al. (2016). A comparison of inferential methods for highly nonlinear state space models in ecology and epidemiology. *Statistical Science 31*(1), 96–118. MR3458595

Fasiolo, M. and S. N. Wood (2015). Approximate methods for dynamic ecological models. *arXiv preprint arXiv:1511.02644*.

Feuerverger, A. (1989). On the empirical saddlepoint approximation. *Biometrika 76*(3), 457–464. MR1040640

Fischer, R., F. Bohn, M. D. de Paula, C. Dislich, J. Groeneveld, A. G. Gutiérrez, M. Kazmierczak, N. Knapp, S. Lehmann, S. Paulick, et al. (2016). Lessons learned from applying a forest gap model to understand ecosystem and carbon dynamics of complex tropical forests. *Ecological Modelling*.

Fukunaga, K. and L. Hostetler (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory 21*(1), 32–40. MR0388638

Gutmann, M. U. and J. Corander (2016). Bayesian optimization for likelihood-free inference of simulator-based statistical models. *The Journal of Machine Learning Research 17*(1), 4256–4302. MR3555016

Hartig, F., J. M. Calabrese, B. Reineking, T. Wiegand, and A. Huth (2011). Statistical inference for stochastic simulation models–theory and application. *Ecology letters 14*(8), 816–827.

Hartig, F., C. Dislich, T. Wiegand, and A. Huth (2014). Technical note: Approximate Bayesian parameterization of a process-based tropical forest model. *Biogeosciences 11*, 1261–1272.

Ionides, E. L., A. Bhadra, Y. Atchadé, and A. King (2011). Iterated filtering. *The Annals of Statistics 39*(3), 1776–1802. MR2850220

Ionides, E. L., C. Bretó, and A. A. King (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences 103*(49), 18438–18443.

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis 97*(10), 2177–2189. MR2301633

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences 100*(26), 15324–15328.

McCullagh, P. (1987). *Tensor methods in statistics*, Volume 161. Chapman and Hall London. MR0907286

Meeds, E. and M. Welling (2014). GPS-ABC: Gaussian process surrogate approximate Bayesian computation. *arXiv preprint arXiv:1401.2838*.

Monti, A. C. and E. Ronchetti (1993). On the relationship between empirical likelihood and empirical saddlepoint approximation for multivariate m-estimators. *Biometrika 80*(2), 329–338. MR1243508

Murray, L. M., A. Lee, and P. E. Jacob (2016). Parallel resampling in the particle filter. *Journal of Computational and Graphical Statistics 25*(3), 789–805. MR3533638

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, 1161–1167. MR1113551

Owen, A. B. (2001). *Empirical likelihood*. CRC press.

Prangle, D. et al. (2017). Adapting the ABC distance function. *Bayesian Analysis 12*(1), 289–309. MR3620131

Price, L. F., C. C. Drovandi, A. Lee, and D. J. Nott (2017). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, 1–11. MR3788296

Rao, C. R. (2009). *Linear statistical inference and its applications*, Volume 22. John Wiley & Sons. MR0221616

Rencher, A. C. and W. F. Christensen (2012). *Methods of multivariate analysis*, Volume 709. John Wiley & Sons. MR2962097

Roberts, A. W. and D. E. Varberg (1973). Convex functions. MR0442824

Ronchetti, E. and A. H. Welsh (1994). Empirical saddlepoint approximations for multivariate M-estimators. *Journal of the Royal Statistical Society. Series B (Methodological) 52*(2), 313–326. MR1281936

Rubio, F. J., A. M. Johansen, et al. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics 7*, 1632–1654. MR3070873

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, Volume 26. CRC press. MR0848134

Spall, J. C. (2005). *Introduction to stochastic search and optimization: estimation, simulation, and control*, Volume 65. John Wiley & Sons. MR1968388

Toni, T., D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface 6*(31), 187–202.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press. MR1652247

Wang, S. (1992). General saddlepoint approximations in the bootstrap. *Statistics & Probability letters 13*(1), 61–66. MR1147640

Wegmann, D., C. Leuenberger, and L. Excoffier (2009). Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics 182*(4), 1207–1218.

Wilkinson, R. (2014). Accelerating ABC methods using Gaussian processes. In *AISTATS*, pp. 1015–1023.

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature 466*(7310), 1102–1104.

Yan, J. et al. (2007). Enjoy the joy of copulas: with a package copula. *Journal of Statistical Software 21*(4), 1–21.