Bayesian Inference Using Synthetic Likelihood: Asymptotics and Adjustments

David T. Frazier, David J. Nott, Christopher Drovandi, and Robert Kohn

1. Introduction

Synthetic likelihood is a popular method used in likelihood-free inference when the likelihood is intractable, but it is possible to simulate from the model for any given parameter value. The method takes a vector summary statistic that is informative about the parameter and assumes it is multivariate normal, estimating the unknown mean and covariance matrix by simulation. Previous research demonstrates that the Bayesian implementation of synthetic likelihood can be more computationally efficient than approximate Bayesian computation, a popular likelihood-free method, in the presence of a high-dimensional summary statistic.

This article makes three contributions. First, it investigates the asymptotic properties of synthetic likelihood when the summary statistic satisfies a central limit theorem. The conditions required for the results are similar to those used by Frazier et al. (2018) in the asymptotic analysis of ABC algorithms, but with an additional assumption controlling the uniform behavior of summary statistic covariance matrices. Under appropriate conditions, the posterior density is asymptotically normal and it quantifies uncertainty accurately, similarly to ABC approaches (Li and Fearnhead 2018a, 2018b; Frazier et al. 2018).

The second contribution demonstrates that a rejection sampling BSL algorithm has a non-negligible acceptance probability when using a “good” proposal density. A similar ABC algorithm has an acceptance probability that goes to zero asymptotically, and in this computational aspect we find that synthetic likelihood performs similarly to regression-adjusted ABC (Li and Fearnhead 2018a, 2018b).

The third contribution considers situations where a parsimonious but misspecified form is assumed for the covariance matrix of the summary statistic, such as a diagonal matrix or a factor model, to speed up the computation. For example, Priddle et al. (2022) show that for a diagonal covariance matrix, the number of simulations need only grow linearly with the summary statistic dimension to control the variance of the synthetic likelihood estimator, as opposed to quadratically for the full covariance matrix. This is especially important for models where simulating the summary statistic is expensive. We use the asymptotic results to motivate sandwich-type variance adjustments to account for the misspecification and implement them in several examples. These adjustments are also potentially useful when the model for the original data is misspecified and we wish to carry out inference for the pseudo-
true parameter value with the data generating density closest to the truth.

For the adjustment methods to be valid, it is important that the summary statistic satisfies a central limit theorem, so that we can make use of the asymptotic normality of the posterior density. This means that these adjustments are not useful for correcting for the effects of violating the normality assumption for the summary statistic. Müller (2013) considers some related methods, although not in the context of synthetic likelihood or likelihood-free inference. Frazier, Robert, and Rousseau (2020) study the consequences of misspecification for ABC approaches to likelihood-free inference.

Wood (2010) introduces the synthetic likelihood and uses it for approximate (non-Bayesian) inference. Price et al. (2018) discuss Bayesian implementations focusing on efficient computational methods. They also show that the synthetic likelihood scales more easily to high-dimensional problems and that computational methods. Price et al. (2018) study the consequences of misspecification for ABC approaches to likelihood-free inference.

Bayesian Synthetic Likelihood

The synthetic likelihood is based on the idea of approximating the likelihood of the summary statistics, as they can only be justified when a true parameter value with the data generating density closest to the unknown \( \theta \).

Like the method of ABC, BSL is most commonly implemented by replacing the observed data \( y \) by a low-dimensional vector of summary statistics. Throughout, the function \( S_n : \mathbb{R}^n \to \mathbb{R}^d, d \geq d_0 \), represents the chosen vector (function) of summary statistics. When confusion is unlikely to result, we abuse notation and let \( S_n = S_n(y) \). For a given model \( p(\theta) \), let \( z = (z_1, \ldots, z_n)^T \) denote data generated under the model \( p(\theta) \), and let \( b(\theta) := E[S_n(z)|\theta] \) and \( \Sigma_n(\theta) := \text{var}[S_n(z)|\theta] \) denote the mean and variance of the summaries calculated under \( p(\theta) \). The map \( \theta \mapsto b(\theta) \) may technically depend on \( n \). However, if the data are independent and identically distributed or weakly dependent, and if \( S_n \) can be written as an average, \( b(\theta) \) will not meaningfully depend on \( n \). As the vast majority of summaries used in BSL satisfy this scenario, it is reasonable to neglect the potential dependence on \( n \).

The synthetic likelihood method approximates the intractable likelihood of \( S_n(z) \) by a normal likelihood. If \( b(\theta) \) and \( \Sigma_n(\theta) \) are known, then the synthetic likelihood is

\[
g_\theta(S_n|\theta) := N\{S_n; b(\theta), \Sigma_n(\theta)\};
\]

here, and below, \( N(\mu, \Sigma) \) denotes a normal distribution with mean \( \mu \) and covariance matrix \( \Sigma \), and \( N(\mu; \Sigma_S, \Sigma_M) \) is its density function evaluated at \( x \).

The idealized BSL posterior is

\[
\pi(\theta|S_n) = \frac{g_\theta(S_n|\theta)\pi(\theta)}{\int_\Theta g_\theta(S_n|\theta)\pi(\theta)d\theta},
\]

which we assume exists for all \( n \). When \( b(\theta) \) and \( \Sigma_n(\theta) \) are known, Markov chain Monte Carlo (MCMC) is used to obtain draws from the target posterior \( \pi(\theta|S_n) \).

While the idealized BSL posterior resembles the quasi-posterior distributions analyzed in, for example, Chernozhukov and Hong (2003) and Bissiri, Holmes, and Walker (2016), outside of simple examples inference based on \( \pi(\theta|S_n) \) is infeasible: \( b(\theta) \) and \( \Sigma_n(\theta) \) can only be analytically calculated if the mean and variance of \( S_n(\cdot) \) are known. Consequently, BSL is generally implemented by replacing \( b(\theta) \) and \( \Sigma_n(\theta) \) with estimates \( \hat{b}_n(\theta) \) and \( \hat{\Sigma}_n(\theta) \). To obtain these estimates, we generate \( m \) independent summary statistics \( \{S_n(z^i)\}_{i=1}^m \), where \( z^i \sim p_\theta \), and take \( \hat{b}_n(\theta) \) as the sample mean of the \( S_n(z^i) \) and \( \hat{\Sigma}_n(\theta) \) as their sample covariance matrix. The notation does not show the dependence of \( \hat{b}_n(\theta) \) and \( \hat{\Sigma}_n(\theta) \) on \( m \), since \( m \) is later taken as a function of \( n \). In practical applications of BSL, the use of variance estimates other than \( \hat{\Sigma}_n(\theta) \) is common (e.g., Ong et al. 2018b; An et al. 2019; Priddle et al. 2022). To cover these and other situations, we take \( \Delta_n(\theta) \) to be a general covariance matrix estimator.

When \( b(\theta) \) and \( \Sigma_n(\theta) \) are replaced by \( \hat{b}_n(\theta) \) and \( \hat{\Sigma}_n(\theta) \), BSL attempts to sample the following target posterior

\[
\hat{\pi}(\theta|S_n) \propto \pi(\theta)\hat{g}_\theta(S_n|\theta),
\]

2. Bayesian Synthetic Likelihood

Let \( y = (y_1, \ldots, y_n)^T \) denote the observed data and define \( p_\theta^{(n)} \) as the true distribution generating \( y \). The model \( p_\theta^{(n)} \) is approximated using a parametric family of models \( \{p_\theta^{(n)} : \theta \in \Theta \subseteq \mathbb{R}^{d_0}\} \), and \( \Pi \) denotes the prior distribution over \( \Theta \), with density \( \pi(\theta) \). We are interested in situations where, due to the complicated nature of the model, the likelihood of \( p_\theta^{(n)} \) is intractable. In such cases, approximate methods such as BSL can be used to conduct inference on the unknown \( \theta \).

As mentioned above, the adjustments for misspecification developed here do not contribute to this literature on robustifying synthetic likelihood inferences to nonnormality of the summary statistics, as they can only be justified when a central limit theorem holds for the summary statistic. Bayesian analyses involving pseudo-likelihoods have been considered in the framework of Laplace-type estimators discussed in Chernozhukov and Hong (2003), but their work does not deal with settings where the likelihood itself must be estimated using Monte Carlo. Forneron and Ng (2018) develop some theory connecting ABC approaches with simulated minimum distance methods widely used in econometrics, and their discussion is also relevant to simulation versions of Laplace-type estimators.
where, for $q_n(\cdot | \theta)$ the density of the simulated summary statistics under $p(\theta)$,

$$
\hat{g}_n(S_n | \theta) := \int N(S_n, \hat{b}_n(\theta), \Delta_n(\theta)) \prod_{i=1}^m q_n(S_n(z^i) | \theta) \ dS_n(z^1) \ldots \ dS_n(z^m),
$$

(2)

with $\hat{g}_n(S_n | \theta)$ the expectation of $N(S_n, \hat{b}_n(\theta), \Delta_n(\theta))$ over the $m$ simulated datasets. Throughout the remainder, we refer to $\hat{\pi}(\theta | S_n)$ as the BSL posterior.

In general, the integration defining $\hat{g}_n(S_n | \theta)$ cannot be performed analytically and $\hat{g}_n(S_n | \theta)$ can be viewed as an intractable (marginal) likelihood. Luckily, an unbiased estimator of $\hat{g}_n(S_n | \theta)$ can be obtained by taking a single set of $m$ independent draws, with $S_n(z^i) \sim q_n(\cdot | \theta)$. Consequently, following arguments in Andrieu and Roberts (2009), a pseudo-marginal algorithm employing an estimator of $\hat{g}_n(S_n | \theta)$ results in sampling from the target posterior $\hat{\pi}(\theta | S_n)$ in (1). However, since $\hat{g}_n(S_n | \theta)$ is a biased estimator of $g_n(S_n | \theta)$, the target BSL posterior $\hat{\pi}(\theta | S_n)$ will differ from the idealized BSL posterior $\pi(\theta | S_n)$. More generally, since $\hat{\pi}(\theta | S_n)$ depends on an intractable (marginal) likelihood, $\hat{g}_n(S_n | \theta)$, it cannot be readily interpreted as a quasi-posterior in the sense of Chernozhukov and Hong (2003) and Bissiri, Holmes, and Walker (2016).

Under idealized, but useful assumptions, Pitt et al. (2012), Doucet et al. (2015) and Sherlock et al. (2015) choose the number of samples $m$ in pseudo-marginal MCMC to optimize the time normalized variance of the posterior mean estimates. They show that a good choice of $m$ occurs (for a given $\theta$) when the variance $\sigma^2(\theta)$ of the log of the likelihood estimator lies between 1 and 3, with a value of 1 suitable for a very good proposal, that is, close to the posterior, and around 3 for an inefficient proposal, for example, a random walk. Deligiannidis, Doucet, and Pitt (2018) propose a correlated pseudo-marginal sampler that tolerates a much greater value of $\sigma^2(\theta)$, and hence a much smaller value of $m$, when the random numbers used to construct the estimates of the likelihood at both the current and proposed values of $\theta$ are correlated; see also Tran et al. (2016) for an alternative construction of a correlated block pseudo-marginal sampler.

Here, the perturbed BSL likelihood is (2) and the log of its estimate is,

$$
-\frac{1}{2} \log |\Delta_n(\theta)| - \frac{1}{2} \left\{ S_n - \hat{b}_n(\theta) \right\}^T \Delta_n^{-1}(\theta) \left\{ S_n - \hat{b}_n(\theta) \right\},
$$

(3)

omitting additive terms not depending on $\theta$. It is straightforward to incorporate either the correlated or block pseudo-marginal approaches into the estimation and show that (3) is bounded in a neighborhood of $\theta_0$ if the eigenvalues of $\Delta_n(\theta)$ are bounded away from zero, suggesting that the variance of the log of the estimate of the synthetic likelihood (3) does not have a high variance in practice. We do not not derive theory for how to select $m$ optimally because that requires taking account of the bias and variance of the synthetic likelihood, which is unavailable in general due to the intractability of the likelihood. However, our empirical work limits $\sigma^2(\theta)$ to lie between 1 and 3, which produces good results. Price et al. (2018) find in their examples that the approximate posterior in (1) depends only weakly on the choice of $m$, and hence they often choose a small value of $m$ for faster computation.

The BSL posterior in (1) is constructed from three separate approximations: (a) the representation of the observed data $y$ by the summaries $S_n(y)$; (b) the approximation of the unknown distribution for the summaries by a Gaussian with unknown mean $b(\theta)$ and covariance $\Sigma_n(\theta)$; (c) the approximation of the unknown mean and covariance by the estimates $\hat{b}_n(\theta)$ and $\Delta_n(\theta)$.

Given the various approximations involved in BSL, it is critical to understand precisely how these approximations impact the resulting inferences on the unknown parameters $\theta$. In practice, understanding how $m$ and $\Delta_n(\theta)$ affect the resulting inferences is particularly important. The larger $m$, the more time consuming is the computation of the BSL posterior. Replacing $\Sigma_n(\theta)$, the covariance of the summaries, by $\Delta_n(\theta)$ means that the posterior may not reliably quantify uncertainty if $\Delta_n(\theta)$ is not carefully chosen. Any theoretical analysis of the BSL posterior is made difficult by the intractability of $\hat{\pi}(\theta)$, and ensures that exploring the finite-sample behavior of the BSL likelihood estimate in (3), and ultimately $\hat{\pi}(\theta | S_n)$, is difficult in general problems. We therefore use asymptotic methods to study the impact of the various approximations within BSL on the resulting inference for $\theta$.

### 3. Asymptotic Behavior of BSL

This section contains several results that disentangle the impact of the previously mentioned approximations used in BSL. These demonstrate that, under regularity conditions, BSL delivers inferences that are just as reliable as other approximate Bayesian methods, such as ABC. Moreover, unlike the commonly applied accept/reject ABC, the acceptance probability obtained by running BSL does not converge to zero as the sample size increases, and is not affected by the number of summaries (assuming they are of fixed dimension, that is, $d = \dim(S_n)$ does not change as $n$ increases).

A Bernstein von-Mises result is first proved and is then used to deduce asymptotic normality of the BSL posterior mean. Using these results, we can demonstrate that valid uncertainty quantification in BSL requires: (a) $m \to \infty$ as $n \to \infty$; (b) the chosen covariance matrix $\Delta_n(\theta)$ used in BSL must be a consistent estimator for the asymptotic variance of the observed summaries $S_n(y)$.

Some notation is now defined to make the results below easier to state and follow. For $x \in \mathbb{R}^d$, $|x|$ denotes the Euclidean norm of $x$. For any matrix $M \in \mathbb{R}^{d \times d}$, we define $|M|$ as the determinant of $M$, and, with some abuse of notation, let $\|M\|$ denote any convenient matrix norm of $M$; the choice of $\| \cdot \|$ is immaterial since we will always be working with matrices of fixed dimension, so that all matrix norms are equivalent. Let $\text{Int}(\theta)$ denote the interior of the set $\theta$. Throughout, $C$ denotes a generic positive constant that can change with each use. For real-valued sequences $\{a_n\}_{n \geq 1}$ and $\{b_n\}_{n \geq 1}$: $a_n \lesssim b_n$ denotes $a_n \leq C b_n$ for some finite $C > 0$ and all $n$ large, $a_n \asymp b_n$ implies $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For $x_n$ a random variable, $x_n = o_p(a_n)$ if $\lim_{n \to \infty} \text{pr}(|x_n/a_n| \geq C) = 0$ for any $C > 0$,
and $x_n = O_P(a_n)$ if for any $C > 0$ there exists a finite $M > 0$ and a finite $n'$ such that, for all $n > n'$, $\text{pr}(\{x_n / a_n \geq M\}) \leq C$. All limits are taken as $n \to \infty$, so that, when there is no confusion, $\lim_n$ denotes $\lim_{n \to \infty}$. The notation $\Rightarrow$ denotes weak convergence. The supplementary materials contains proofs of all stated results.

### 3.1. Asymptotic Behavior of the BSL Posterior

This section establishes the asymptotic behavior of the BSL posterior $\hat{\pi}(\theta | x_n)$ in Equation (1). We do not assume that $\Delta_n(\theta)$ is a consistent estimator of $\Sigma_n(\theta)$ to allow the synthetic likelihood covariance to be “misspecified.” The following regularity conditions are assumed on $S_n, b(\theta)$ and $\Delta_n(\theta)$.

**Assumption 1.** There exists a sequence of positive real numbers $\nu_n$ diverging to $0$ and a vector $b_0 \in \mathbb{R}^d$, $d \geq d_0$, such that $V_0 := \lim_n \text{var}(\nu_n S_n - b_0)$ exists and $\nu_n (S_n - b_0) \Rightarrow N(0, V_0)$ under $P_0^{(\theta)}$.

**Assumption 2.** (i) The map $\theta \mapsto b(\theta)$ is continuous, and there exists a unique $\theta_0 \in \text{Int}(\Theta)$, such that $b(\theta_0) = b_0$; (ii) for some $\delta > 0$, and all $\|\theta - \theta_0\| \leq \delta$, the Jacobian $\nabla b(\theta)$ exists and is continuous, and $V b(\theta_0)$ has full column rank $d_0$.

**Assumption 3.** The following conditions are satisfied for some $\delta > 0$: (i) for $n$ large enough, the matrix $\nabla^2 \Delta_n(\theta)$ is positive semidefinite and symmetric uniformly over $\Theta$, and positive-definite for all $\|\theta - \theta_0\| \leq \delta$; (ii) there exists some matrix $\Delta(\theta)$, positive semidefinite and symmetric uniformly over $\Theta$, and such that $\sup_{\theta \in \Theta} \|\nabla^2 \Delta_n(\theta) - \Delta(\theta)\| = o_p(1)$, and, for all $\|\theta - \theta_0\| \leq \delta$, $\Delta(\theta)$ is continuous and positive-definite; (iii) for any $\epsilon > 0$, $\sup_{\|\theta - \theta_0\| \geq \epsilon} - \{b(\theta) - b_0\}^T \Delta(\theta)^{-1}\{b(\theta) - b_0\} < 0$.

**Assumption 4.** For $\theta_0$ defined in Assumption 2, $\pi(\theta_0) > 0$, and $\pi(\cdot)$ is continuous on $\Theta$. For some $p > 0$, and $n$ large enough, $\int_\Theta \|\theta\|^p \pi(\theta) d\theta < \infty$ and $\int_\Theta \|\nabla^2 \Delta_n(\theta)\|^{-1/2} \|\theta\|^p \pi(\theta) d\theta < \infty$.

**Assumption 5.** There exists a function $k : \Theta \to \mathbb{R}$, such that: (i) for all $\alpha \in \mathbb{R}^d, E(\exp [\alpha^T v_n (S_n(z) - b(\theta_0))] \leq \exp [\frac{1}{2} \|\alpha\|^2 k(\theta_0)/2]$; (ii) there exists a constant $k \geq 0$ such that $k(\theta) \leq \|\theta\|^p$; (iii) for all $n$ large enough, $\sup_{\theta \in \Theta} [k(\theta)]^{1/2} \|\nabla^2 \Delta_n(\theta)\|^{-1/2} \|\cdot\| < \infty$.

These assumptions are similar to those used to prove Bernstein–von Mises results in ABC (Frazier et al. 2018; Li and Fearnhead 2018a). In particular, Assumption 1 requires that the observed summaries satisfy a central limit theorem. Assumption 2 ensures that, over $\Theta$, the summaries $S_n(z)$ have a well-behaved limit $b(\theta)$ that is continuous over $\Theta$, can identify $\theta_0$, and whose derivative has full column rank at $\theta_0$. Assumption 2 does not require that $P_0^{(\theta)}$ corresponds to $P_0^{(\theta_0)}$, so that the model can be misspecified, but instead requires the weaker condition that there exists a unique value $\theta_0 \in \Theta$ under which $b(\theta_0) = b_0$, referred to subsequently as the “true” parameter value.

Variants of Assumption 4 are commonly encountered in the literature on Bayesian asymptotics. In addition to the continuity of $\pi(\theta)$, Assumption 4 requires the existence of a certain prior moment. This condition is slightly stronger than the prior moment condition needed in the standard case. The need to strengthen this assumption comes from the possibility that the matrix $\Delta_n(\theta)$ may be singular far away from $\theta_0$. As such, in order to ensure the BSL posterior is well-behaved, we require that the prior has thin enough tails in the region where $\Delta_n(\theta)$ is singular, so that the potential singularity of $\Delta_n(\theta)$ does not impact posterior concentration. When $\Delta(\theta)$ in Assumption 3 is positive-definite, uniformly over $\Theta$, this latter condition can be replaced by the standard assumption that $\int_\Theta \|\theta\|^p \pi(\theta) d\theta < \infty$ for some $p > 0$.

Globally speaking, Assumption 5 requires that the simulated summaries are sub-Gaussian. Intuitively, this condition requires that the simulated summaries have an exponential moment, and is similar to certain conditions employed by Frazier et al. (2018) for ABC. We refer to Vershynin (2018) for a textbook treatment on the properties of sub-Gaussian random variables, and note that many different classes of random variables satisfy this assumption. Without further conditions on the number of model simulations $m$, some version of this assumption seems necessary to ensure that the BSL posterior exists, since $\hat{\pi}(\Theta, x_n)$ is, by definition, an expectation—with respect to the distribution of the simulated summaries—of an exponentiated quadratic form.

The key difference between the current assumptions and those used in the theoretical analysis of ABC is that in BSL the behavior of the quadratic form $\|\Delta_n^{-1/2}(\theta) (b(\theta) - S_n)\|^2$ determines the behavior of the synthetic likelihood, and needs to be controlled. Assumption 3(i) requires that, for $n$ large enough, the matrix in this quadratic form is positive-definite for any $\theta$ sufficiently close to $\theta_0$, while Assumption 3(ii) requires that $\nabla^2 \Delta_n(\theta)$ converges uniformly to $\Delta(\theta)$, which is continuous and positive-definite for all $\theta$ sufficiently close to $\theta_0$. Assumption 3(ii) does not require $\Delta(\theta)$ to be positive-definite uniformly over $\Theta$, and thus it is unnecessary for it to be invertible far from $\theta_0$. This implies that the quadratic form $\|\Delta_n^{-1/2}(\theta) (b(\theta) - S_n)\|^2$ may not be continuous (or finite) uniformly over $\Theta$. In such situations, it is necessary to maintain the additional identification assumption given in Assumption 3(iii). However, if $\Delta(\theta)$ is continuous over $\Theta$ this identification assumption is automatically satisfied.

Assumptions 1–5 are sufficient to deduce a Bernstein von-Mises result for the BSL posterior. To state this result, define the local parameter $t := v_n (\theta - \theta_0) - W_0^{-1} Z_n$,

where $Z_n := \nabla b(\theta_0)^T \Delta(\theta_0)^{-1} v_n (S_n - b(\theta_0))$, $W_0 := \{\nabla b(\theta_0)^T \Delta(\theta_0)^{-1} \nabla b(\theta_0)\}$, and denote the BSL posterior for $t$ as $\hat{\pi}(t | S_n) := \hat{\pi} \left( t + t/v_n + W_0^{-1} Z_n v_n | S_n \right) / v_n^{\theta_0}$.

The support of $t$ is $T_n := \{v_n (\theta - \theta_0) - W_0^{-1} Z_n : \theta \in \Theta\}$, which is a scaled and shifted translation of $\Theta$. The following result states that the total variation distance between $\hat{\pi}(t | S_n)$ and $N(t; 0, W_0^{-1})$ converges to zero in probability. It also demonstrates that the covariance of the Gaussian density to which $\hat{\pi}(t | S_n)$ converges depends on the variance estimator $\Delta_n(\theta)$ used in BSL.
Theorem 1. If Assumptions 1–5 are satisfied, with $p \geq \kappa$ in Assumption 4, and if $m = m(n) \to \infty$ as $n \to \infty$, then

$$\int_{T_n} |\hat{\pi}(t|S_n) - N\{t; 0, W^{-1}_0\}| \, dt = o_p(1).$$

Furthermore, for any $0 < \gamma \leq 2$, if Assumption 4 is satisfied with $p \geq \gamma + \kappa$, then

$$\int_{T_n} \|t\|^{\gamma} \left|\hat{\pi}(t|S_n) - N\{t; 0, W^{-1}_0\}\right| \, dt = o_p(1).$$

The second result in Theorem 1 demonstrates that, under moment assumptions on the prior, the mean difference between the BSL posterior $\hat{\pi}(t|S_n)$ and $N\{t; 0, W^{-1}_0\}$ converges to zero in probability. Using this result, we can demonstrate that the BSL posterior mean $\hat{\theta}_n := \int_{\Theta} \theta \hat{\pi}(\theta|S_n) \, d\theta$ is asymptotically Gaussian with a covariance matrix that depends on the version of $\Delta_n(\theta)$ used in the synthetic likelihood.

Corollary 1. If the Assumptions in Theorem 1 are satisfied, then for $m \to \infty$ as $n \to \infty$,

$$v_n(\hat{\theta}_n - \theta_0) \Rightarrow N\left[0, W^{-1}_0 \left\{\nabla b(\theta_0)^T \Delta(\theta_0)^{-1} V_0 \Delta(\theta_0)^{-1} \nabla b(\theta_0)\right\} W^{-1}_0\right],$$

under $P^{(n)}$.

Remark 1. The above results only require weak conditions on the number of simulated datasets, $m$, and are satisfied for any $m \in \mathcal{O}(n^2)$, with $C > 0$, $\gamma > 0$, and $|x|$ denoting the integer floor of $x$. Therefore, Theorem 1 and Corollary 1 demonstrate that the choice of $m$ does not strongly impact the resulting inference on $\theta$ and its choice should be driven by computational considerations. We note that this requirement is in contrast to ABC, where the choice of tuning parameters, that is, the tolerance, significantly impacts both the theoretical behavior of ABC and the practical (computing) behavior of ABC algorithms. However, this lack of dependence on tuning parameters comes at the cost of requiring that a version of Assumptions 3 and 5 are satisfied. ABC requires no condition similar to Assumption 3, while Assumption 5 is stronger than the tail conditions on the summaries required for the ABC posterior to be asymptotically Gaussian.

Remark 2. Theorem 1 and Corollary 1 demonstrate the tradeoff between using a parsimonious choice for $\Delta_n(\theta)$, leading to faster computation, and a posterior that correctly quantifies uncertainty. BSL credible sets provide valid uncertainty quantification, in the sense that they have the correct level of asymptotic coverage, when

$$\int_{T_n} t \hat{\pi}(t|S_n) \, dt = W^{-1}_0 \nabla b(\theta_0)^T \Delta(\theta_0)^{-1} V_0 \Delta(\theta_0)^{-1} W^{-1}_0 + o_p(1).$$

However, the second part of Theorem 1 implies that

$$\int_{T_n} t \hat{\pi}(t|S_n) \, dt = W^{-1}_0 + o_p(1) = [\nabla b(\theta_0)^T \Delta(\theta_0)^{-1} \nabla b(\theta_0)]^{-1} + o_p(1),$$

so that a sufficient condition for the BSL posterior to correctly quantify uncertainty is that

$$\Delta(\theta_0) = V_0.$$ (4)

Satisfying Equation (4) generally requires using the more computationally intensive variance estimator $\hat{\Sigma}_n(\theta)$, and that the variance model is “correctly specified”; here, correctly specified means $\theta_0$ satisfies $b(\theta_0) = b_0$ and $\theta_0$ also satisfies Equation (4), and where we note that the latter condition is not implied by Assumptions 1–5. While a sufficient condition for (4) is that $P^{(n)}_\theta = P^{(n)}_0$ for some $\theta_0 \in \Theta$, this condition is not necessary in general. Given the computational costs associated with using $\hat{\Sigma}_n(\theta)$ when the summaries are high-dimensional, Section 4 proposes an adjustment approach to BSL that allows the use of the simpler, possibly misspecified, variance estimator $\Delta_n(\theta)$, but which also yields a posterior that has valid uncertainty quantification.

Remark 3. In contrast to ABC point estimators, Corollary 1 demonstrates that BSL point estimators are generally asymptotically inefficient. It is known that $\left\{\nabla b(\theta_0)^T V_0^{-1} \nabla b(\theta_0)^T\right\}^{-1}$ is the smallest achievable asymptotic variance for any $\nu_n$-consistent and asymptotically normal estimator of $\theta_0$ based on parametric classes of models $\{P^{(n)}_\theta : \theta \in \Theta\}$ and conditional on the summary statistics $S_n$; see, for example, Li and Fearnhead (2018b). We also have that

$$W^{-1}_0 \left\{\nabla b(\theta_0)^T \Delta(\theta_0)^{-1} V_0 \Delta(\theta_0)^{-1} \nabla b(\theta_0)\right\} \leq \left\{\nabla b(\theta_0)^T V_0^{-1} \nabla b(\theta_0)^T\right\}^{-1};$$

where for square matrices $A, B$, $A \succeq B$ means that $A - B$ is positive semidefinite. Given this, the BSL posterior mean $\hat{\theta}_n$ is asymptotically efficient only when Equation (4) is satisfied. In this case, BSL simultaneously delivers efficient point estimators and asymptotically correct uncertainty quantification.

Remark 4. The BSL posterior $\hat{\pi}(\theta|S_n)$ depends on the intractable (marginal) likelihood $\hat{\pi}_n(S_n|\theta)$, which is an expectation of a normal density with estimated mean and variance (see Equations (1)–(2)) and is a biased estimator of the normal density $N(S_n; b(\theta), \Sigma_n(\theta))$. Consequently, existing large sample results on the behavior or quasi-posterior distributions cannot be used to directly ascertain the behavior of the BSL posterior. Indeed, in contrast to the case of quasi-posteriors, the proof of Theorem 1 shows that to control the behavior of the BSL posterior, we must control the total variation norm of the expected difference (with respect to $(z^1, \ldots, z^m)$) between the estimated and idealized synthetic likelihoods. We elaborate on this point in Section B.3 of the supplementary materials, give further details regarding the proof strategy used to obtain our main results, and compare our strategy with others that have featured in the likelihood-free and quasi-posterior literatures.

3.2. Computational Efficiency

Li and Fearnhead (2018a, 2018b) discuss the computational efficiency of vanilla and regression-adjusted ABC algorithms using a rejection sampling method based on a “good” proposal density $q_n(\theta)$. They show that regression-adjusted ABC yields
asymptotically correct uncertainty quantification, that is, credible sets with the correct level of frequentist coverage, and an asymptotically nonzero acceptance rate, while vanilla ABC can only accomplish one or the other.

This section shows that BSL can deliver correct uncertainty quantification and an asymptotically nonzero acceptance rate, if the number of simulated datasets used in the synthetic likelihood tends to infinity with the sample size. We follow Li and Fearnhead (2018a) and consider implementing synthetic likelihood using a rejection sampling algorithm based on the proposal \( q_n(\theta) \) analogous to the one they consider for ABC. Following Assumption 3(i), there exists a uniform upper bound of the form \( C v_n^2 \) for some \( 0 < C < \infty \) locally in a neighborhood of \( \theta_0 \) on \( N \{ S_n, b(\theta), \Delta_n(\theta) \} \) for \( n \) large enough; an asymptotically valid rejection sampler then proceeds as follows.

**Rejection sampling BSL algorithm:**

1. Draw \( \theta' \sim q_n(\theta) \)
2. Accept \( \theta' \) with probability
   \[
   (C v_n^2)^{-1} \sup_{\theta} q_n(\theta) \left\{ S_n, b_n(\theta'), \Delta_n(\theta') \right\}.
   \]

An accepted value from this sampling scheme is a draw from the density proportional to \( q_n(\theta) \sup_{\theta} q_n(\theta) \). Similarly to the analogous ABC scheme considered in Li and Fearnhead (2018a), samples from this rejection sampler can be reweighted with importance weights proportional to \( \pi(\theta')/q_n(\theta') \) to recover draws from \( \pi(\theta|S_n) \propto \pi(\theta)\hat{g}_n(S_n|\theta) \).

We choose the proposal density \( q_n(\theta) \) to be from the location-scale family \( \mu_n + \sigma_n X \), where \( X \) is a \( d_\theta \)-dimensional random variable such that \( X \sim q(\cdot), E_q[X] = 0 \) and \( E_q[||X||^2] < \infty \). The sequences \( \mu_n \) and \( \sigma_n \) depend on \( n \) and satisfy the following assumption.

**Assumption 6.** (i) There exists a positive constant \( C \), such that \( 0 < \sup_q q(x) \leq C < \infty \); (ii) the sequence \( \sigma_n > 0 \), for all \( n \geq 1 \), satisfies \( \sigma_0 = o(1) \), and \( v_n \sigma_n \rightarrow c_\sigma \), for some positive constant \( c_\sigma \); (iii) the sequence \( \mu_n \) satisfies \( s_n^{-1}(\mu_n - \theta_0) = O(1) \); (iv) for \( h_n(\theta) = q_n(\theta)/\pi(\theta) \), lim \( \sup_{n \rightarrow \infty} \int h_n^2(\theta) \pi(\theta|S_n) d\theta < \infty \).

**Remark 5.** Assumption 6 formalizes the conditions required of the proposal density and are similar to those required in Li and Fearnhead (2018a). It is satisfied if the proposal density \( q_n(\theta) \) is built from \( v_n \)-consistent estimators of \( \theta_0 \), such as those based on pilot runs.

Following Li and Fearnhead (2018a), the acceptance probability associated with this rejection sampling BSL algorithm is given by

\[
\tilde{\alpha}_n := (C v_n^2)^{-1} \int q_n(\theta) \hat{g}_n(S_n|\theta) d\theta.
\]

We measure the computational efficiency of the rejection sampling BSL algorithm via the behavior of \( \tilde{\alpha}_n \). If \( \tilde{\alpha}_n \) is asymptotically nonzero, then by Theorem 1, and under the restriction in (4), implementing a rejection-based BSL approach can yield a posterior that has credible sets with the correct level of frequentist coverage and computational properties that are similar to those of regression-adjusted ABC.

Theorem 2 describes the asymptotic behavior of \( \tilde{\alpha}_n \) using the proposal density given in Assumption 6. The result uses the following definition: for a random variable \( x_n \), we write \( x_n = \mathbb{E}_p(V_n) \) if there exist constants \( 0 < c \leq C < \infty \) such that \( \lim_n P(c < |x_n/V_n| < C) = 1 \).

**Theorem 2.** If Assumptions 1–6 are satisfied and if \( \int k(\theta)^2 \pi(\theta|S_n) d\theta < \infty \), then for \( m \rightarrow \infty \) as \( n \rightarrow \infty \), \( \tilde{\alpha}_n = \mathbb{E}_p(1 + O_p(1/m)) \).

While Theorem 2 holds for all choices of \( \Delta_n(\theta) \) satisfying Assumption 3, taking \( \Delta_n(\theta) = \Sigma_n(\theta) \) implies that the resulting BSL posterior yields credible sets with the appropriate level of frequentist coverage and that the rejection-based algorithm has a non-negligible acceptance rate asymptotically. Therefore, the result in Theorem 2 is a BSL version of Theorem 2 in Li and Fearnhead (2018a), demonstrating a similar result, under particular choices of the tolerance sequence, for regression-adjusted ABC.

In contrast to BSL, in the regime where rejection-based ABC correctly quantifies uncertainty, the ABC acceptance probability decays to zero at rate \( o(1/d_m^2) \) (Corollary 1, Frazier et al. 2018). However, each ABC sample requires only one summary statistic simulation, whereas each BSL sample requires \( m \), with \( m \rightarrow \infty \) as \( n \rightarrow \infty \), and possibly very slowly. To account for the extra simulation cost in BSL, one could divide the BSL acceptance probability \( \tilde{\alpha}_n \) by \( m \) to obtain a “per-simulation acceptance rate.” However, for any reasonable choice of \( m \), the per-simulation acceptance rate \( \tilde{\alpha}_n/m \) converges to zero much more slowly than the rejection-based ABC acceptance probability. Consequently, even though BSL must simulate \( m \) datasets, it is still much more efficient than standard ABC.

The example in Section 3 of Price et al. (2018) compares rejection ABC and a rejection version of synthetic likelihood, where the model is normal and \( \Sigma_n(\theta) \) is constant and does not need to be estimated. They find that with the prior as the proposal, ABC is more efficient when \( d = 1 \), equally efficient when \( d = 2 \), but less efficient than synthetic likelihood when \( d > 2 \). The essence of the example is that the sampling variability in estimating \( b(\theta) \) can be equated with the effect of a Gaussian kernel in their toy normal model for a certain relationship between \( \epsilon \) and \( m \). The discussion above suggests that in general models, and with a good proposal, in large samples the synthetic likelihood is preferable to the vanilla ABC algorithm no matter the dimension of the summary statistic. However, this greater computational efficiency is only achieved through the strong tail assumption on the summaries.

### 4. Adjustments for Misspecification

From the discussion in Remark 2, if BSL uses a misspecified estimator for the variance for the summaries, in the sense that Equation (4) does not hold, then the BSL posterior gives invalid uncertainty quantification. This section outlines one approach for adjusting inferences to account for this form of misspecification when Assumption 2 is satisfied, although, there are other ways to do so. Suppose \( \hat{\theta} = q = 1, \ldots, Q \) is an approximate sample from \( \hat{\pi}(\theta|S_n) \), obtained by MCMC for example. Let \( \tilde{\theta}_n \) denote the synthetic likelihood posterior mean, let \( \hat{\Gamma} \) denote the synthetic likelihood posterior covariance, and write \( \tilde{\theta} \) and \( \hat{\Gamma} \) for their sample estimates based on \( \hat{\theta} \), \( q = 1, \ldots, Q \). Consider the
adjusted sample
\[ \theta^{A,q} = \hat{\theta} + \Gamma \tilde{\Omega}^{1/2} \gamma^{-1/2}(\theta^q - \hat{\theta}), \]

where \( q = 1, \ldots, Q \), \( \tilde{\Omega} \) is an estimate of \( \text{var} \{ \nabla_\theta \log g_n(S_n|\theta) \} \); the estimate of \( \Omega \) is discussed below. We propose using (5) as an approximate sample from the posterior, which is similar to the original sample when the model is correctly specified, but gives asymptotically valid frequentist inference about the pseudo-true parameter value when the model is misspecified. We note that the value of \( m \) used in the adjustment to obtain \( g_n(S_n|\theta) \) need not be the same as the one used in the BSL sampling using a misspecified estimator for the variance. We suggest choosing \( m \) for the adjustment as large as feasible within the computational budget available to reduce the synthetic likelihood variability, making it easier to estimate the quantities required in the above adjustment. Note that the adjustment typically requires orders of magnitude fewer model simulations than MCMC BSL and so a larger value of \( m \) for the adjustment is affordable.

The motivation for (5) is that if \( \theta^q \) is approximately drawn from the normal distribution \( N(\hat{\theta}, \Gamma) \), then \( \theta^{A,q} \) is approximately drawn from \( N(\hat{\theta}, \tilde{\Omega} \Gamma) \). The results of Corollary 1 imply that if \( \tilde{\Omega} \approx \text{var} \{ \nabla_\theta \log g_n(S_n|\theta) \} \) and \( \Gamma \) is approximately the inverse negative Hessian of \( \log g(S_n|\theta) \) at \( \theta_0 \), then the covariance matrix of the adjusted samples is approximately that of the sampling distribution of the BSL posterior mean, giving approximate frequentist validity to posterior credible intervals based on the adjusted posterior samples. We now suggest two ways to obtain \( \tilde{\Omega} \). The first is suitable if the model assumed for \( y \) is true, but the covariance matrix \( \lim_n \nabla_\theta \log g(S_n|\theta) \neq V_0 \), which we refer to as misspecification of the working covariance matrix. The second way is suitable when the models for both \( y \) and the working covariance matrix may be misspecified, but Assumption 2 holds.

4.1. Estimating \( \text{var} \{ \nabla_\theta \log g_n(S_n|\theta_0) \} \) when the Model for \( y \) is Correct

Algorithm 1: Estimating \( \tilde{\Omega} \) when the model for \( y \) is correct.

1. For \( j = 1, \ldots, J \), sample \( y^{(j)} \) with replacement to get a bootstrap sample \( y^{(j)} \) with corresponding summary \( S^{(j)} \).

If the data is dependent it may still be possible to use the bootstrap (Kreiss and Paparoditis 2011); however, the implementation details are model dependent.

4.3. What the Adjustments Can and Cannot Do

The adjustments suggested above are intended to achieve asymptotically valid frequentist inference when the consistency in (4) is not satisfied, that is, when \( \lim_n \Delta_n(\theta) \neq V_0 \), or when the model for \( y \) is misspecified, but \( S_n \) still satisfies a central limit theorem. The adjustment will not recover the posterior distribution that is obtained when the model is correctly specified. Asymptotically valid frequentist estimation based on the synthetic likelihood posterior mean for the misspecified synthetic likelihood is frequentist inference based on a point estimator of \( \theta \) that is generally less efficient than in the correctly specified case. Matching posterior uncertainty after adjustment to the sampling variability of such an estimator does not recover the posterior uncertainty from the correctly specified situation.

5. Examples

5.1. Toy Example

Suppose that \( y_1, \ldots, y_n \) are independent observations from a negative binomial distribution \( NB(5, 0.5) \) so they have mean 5 and variance 10. We model the \( y_i \) as independent and coming from a Poisson(\( \theta \)) distribution and act as if the likelihood is intractable, basing inference on the sample mean \( \bar{y} \) as the summary statistic. The pseudo-true parameter value \( \theta_0 \) is 5, since this is the parameter value for which the summary statistic mean matches the corresponding mean for the true data generating process.

Under the Poisson model, the synthetic likelihood has \( b(\theta) = \theta \) and \( \Delta_n(\theta) = \theta / n \). We consider a simulated dataset with \( n = 20 \), and deliberately misspecify the variance model in the synthetic likelihood under the Poisson model as \( \Delta_n(\theta) = \theta / (2n) \). As noted previously, the deliberate misspecification of \( \text{var}(S_n|\theta) \) may be of interest in problems with a high-dimensional \( S_n \) as a way of reducing the number of simulated summaries needed to estimate \( \text{var}(S_n|\theta) \) with reasonable precision; for example, we might assume \( \text{var}(S_n|\theta) \) is diagonal or based on a factor model.

Figure 1 shows the estimated posterior densities obtained using a number of different approaches, when the prior for \( \theta \) is Gamma(2, 0.5). The narrowest green density is obtained from the synthetic likelihood with a misspecified variance. This density is obtained using 50,000 iterations of a Metropolis-Hastings MCMC algorithm with a normal random walk proposal. The red density is the exact posterior assuming the Poisson likelihood is correct, which is Gamma(2 + \( n \bar{y} \), 0.5 + \( n \)). The purple kernel density estimate is based on the adjusted synthetic likelihood samples; it uses the method of Section 4.1 for the adjustment in which the \( y \) model is assumed correct but the working covariance matrix is misspecified. The figure shows
that the adjustment gives a result very close to the exact posterior under an assumed Poisson model. Finally, the light blue section of the synthetic likelihood, uses the method of Section 4.2 based on the posterior under an assumed Poisson model. This posterior is more dispersed than the one obtained under the Poisson assumption, since the negative binomial generating density is overdispersed relative to the Poisson, and hence the observed \( \hat{y} \) is less informative about the pseudo-true parameter value than implied by the Poisson model.

### 5.2. Examples with a High-Dimensional Summary Statistic

This section explores the efficacy of the adjustment method when using a misspecified covariance in the presence of a high-dimensional summary statistic \( S \). All the examples below use the Warton (2008) shrinkage estimator to reduce the number of simulations required to obtain a stable covariance matrix estimate in the synthetic likelihood. Based on \( m \) independent model simulations the covariance matrix estimate is

\[
\hat{\Sigma}_\gamma = \hat{D}^{1/2} \left\{ \gamma \hat{C} + (1 - \gamma) I \right\} \hat{D}^{1/2},
\]

where \( \hat{C} \) is the sample correlation matrix, \( \hat{D} \) is the diagonal matrix of component sample variances, and \( \gamma \in [0, 1] \) is a shrinkage parameter. The matrix \( \hat{\Sigma}_\gamma \) is nonsingular if \( \gamma < 1 \), even if \( m \) is less than the number of the observations. This estimator shrinks the sample correlation matrix toward the identity. When \( \gamma = 1 \) (resp. \( \gamma = 0 \)) there is no shrinkage (resp. a diagonal covariance matrix is produced). We choose \( \gamma \) to require only 1/10 of the simulations required by the standard synthetic likelihood for Bayesian inference. We are interested in the shrinkage effect on the synthetic likelihood approximation and whether our methods can offer a useful adjustment. Heavy shrinkage is used to stabilize covariance estimation in the synthetic likelihood; so, the shrinkage estimator can be thought of as specifying \( \Delta_n(\theta) \).

To perform the adjustment, it is necessary to approximate the derivative of the synthetic log-likelihood, with shrinkage applied, at a point estimate of the parameter; we take this point as the estimated posterior mean \( \hat{\theta} \) of the BSL approximation. A computationally efficient approach for estimating these derivatives uses Gaussian process emulation of the approximate log-likelihood surface based on a precomputed training sample. The training sample is constructed around \( \hat{\theta} \), because this is the only value of \( \theta \) for which the approximate derivative is required. We sample \( B \) values using Latin hypercube sampling from the hypercube defined by \([\hat{\theta}_k - \delta_k, \hat{\theta}_k + \delta_k] \), where \( \hat{\theta}_k \) denotes the \( k \)th component of \( \hat{\theta} \), and take \( \delta_k \) as the approximate posterior standard deviation of \( \hat{\theta}_k \); see McKay, Beckman, and Conover (1979) for details on Latin hypercube sampling. Denote the collection of training data as \( T = \{\theta^b, \mu^b, \Sigma^b\}_{b=1}^B \), where \( \theta^b \) is the \( b \)th training sample and \( \mu^b \) and \( \Sigma^b \) are the corresponding estimated mean and covariance of the synthetic likelihood from the \( m \) model simulations, respectively. This training sample is stored and recycled for each simulated dataset generated from \( \hat{\theta} \) that needs to be processed in the adjustment method, which is now described in more detail.

For a simulated statistic \( S^{(i)} \) generated from the model at \( \hat{\theta} \), the shrinkage synthetic log-likelihood is rapidly computed at each \( \theta^b \) in the training data \( T \) using the prestored information, denoted as \( l^b = l(\theta^b; S^{(i)}) \). A Gaussian process regression model based on the collection \( \{\theta^b, l^b, \mu^b, \Sigma^b\}_{b=1}^B \) is then fitted with \( l^b \) as the response and \( \theta^b \) as the predictor. We use a zero-mean Gaussian process with squared exponential covariance function having different length scales for different components of \( \theta \) and then approximate the gradient of \( \log g_{\theta}(S^{(i)}|\hat{\theta}) \) by computing the derivative of the smooth predicted mean function of the Gaussian process at \( \hat{\theta} \). We can show that this is equivalent to considering the bivariate Gaussian process of the original process and its derivative, and performing prediction for the derivative value. The derivative is estimated using a finite difference approximation because it is simpler than computing the estimate explicitly. The matrix \( \hat{\Sigma} \) is constructed using \( B = 200 \) training samples and \( J = 200 \) datasets. Both examples below use 20,000 iterations of MCMC for standard and shrinkage BSL with a multivariate normal random walk proposal. In each case, the covariance of the random walk was set based on an approximate posterior covariance obtained by pilot MCMC runs.

#### Moving Average Example

We consider the second order moving average model (MA(2)):

\[
y_t = z_t + \theta_1 z_{t-1} + \theta_2 z_{t-2},
\]

for \( t = 1, \ldots, n \), where \( z_t \sim N(0, 1) \), \( t = -1, \ldots, n \), and \( n \) is the number of observations in the time series. To ensure
identifiability of the MA(2) model, the space $\Theta$ is constrained as $-1 < \theta_2 < 1$, $\theta_1 + \theta_2 > -1$, $\theta_1 - \theta_2 < 1$ and we specify a uniform prior over this region. The density of the observations from an MA(2) model is multivariate normal, with $\text{var}(y_t) = 1 + \theta_1^2 + \theta_2^2$, $\text{cov}(y_t, y_{t-1}) = \theta_1 + \theta_2$, $\text{cov}(y_t, y_{t-2}) = \theta_2$, with all other covariances equal to 0. The coverage assessment is based on 100 simulated datasets from the model with true parameters $\theta_1 = 0.6$ and $\theta_2 = 0.2$. Here, we consider a reasonably large sample size of $n = 10^4$.

This example uses the first 20 autocovariances as the summary statistic. The autocovariances are a reasonable choice here as they are informative about the parameters and satisfy a central limit theorem (Hannan 1976).

To compare with BSL, we use ABC with a Gaussian weighting kernel having covariance $\epsilon V$, where $V$ is a positive-definite matrix. To favor the ABC method, $V$ is set as the covariance matrix of the summary statistic obtained via many simulations at the true parameter value. This ABC likelihood corresponds to using the Mahalanobis distance function with a Gaussian weighting kernel. We also consider BSL with a diagonal covariance, and the corresponding adjustment described in Section 4.

To sample from the approximate posterior distributions for each method and dataset, importance sampling with a Gaussian proposal is used with a mean given by the approximate posterior mean and a covariance that is twice the approximate posterior covariance. We treat this as the "good" proposal distribution for posterior inference. The initial approximations of the (approximate) posteriors are obtained from pilot runs.

For BSL, we use 10,000 importance samples and consider $m = 100, 200, 500, 1000$ for estimating the synthetic likelihood. Table 1 reports the mean and minimum effective sample size (ESS) of the importance sampling approximations (Kong 1992) over the 100 datasets. It shows that for standard BSL with $m = 100$ the minimum ESS is small, suggesting this is close to the smallest value of $m$ that can be considered to ensure the results are not dominated by Monte Carlo error. For a given $m$, the ESS values are larger when using a diagonal covariance matrix, demonstrating the computational benefit over estimating a full covariance matrix in standard BSL. For the BSL adjustment approach, the initial sample before adjustment consists of a resample of size 1000 from the relevant diagonal BSL importance sampling approximation to avoid having to work with a weighted sample.

We use 10 million importance samples for ABC; twice as many model simulations compared to BSL with $m = 500$. For each dataset, $\epsilon$ is selected so that the ESS is around 1000, to reduce $\epsilon$ as much as possible, while ensuring that the results are robust to Monte Carlo error. To reduce storage, a resample of size 1000 is taken from the ABC importance sampling approximation to produce the final ABC approximation. We also apply the local regression adjustment of Beaumont, Zhang, and Balding (2002) to the ABC samples for each dataset.

Table 1 presents the estimated marginal coverage rates for $\theta_1$, $\theta_2$ marginally, and the joint coverage for $(\theta_1, \theta_2)$, for nominal coverages of 95%, 90%, and 80% using kernel density estimates. Total variation distances to the true posterior are also estimated using kernel density estimation, with samples from the true posterior drawn using MCMC. The densities are estimated from 1000 samples, performing resampling from the importance sampling approximations when required to avoid dealing with a weighted sample.

It is evident that standard BSL produces reasonable coverage rates, with some under-coverage at the 80% nominal rate; $m$ seems to have negligible effect on the estimated coverage. BSL with a diagonal covariance produces gross over-coverage for $\theta_1$. Interestingly, despite the over-coverage for $\theta_1$, there is under-coverage at the 95% and 90% nominal rates for the joint confidence regions for $\theta_1$ and $\theta_2$, due to the incorrectly estimated dependence structure based on the misspecified covariance. In contrast, the adjusted BSL results produce accurate coverage rates for the marginals and the joint.

The ABC method produces substantial over-coverage. ABC with regression adjustment produces more accurate coverage rates, although some over-coverage remains in general. Table 1 also includes some approximate computing times for each of the approaches. Each run for each method is performed on a single core of an Intel Xeon Silver 4116 Processor.

In Section B.1 of the supplementary materials we further explore how BSL scales as the sample size and number of summary statistics increases. In the vast majority of cases, we find
that standard BSL gives the closest approximation to the true posterior across the different BSL varieties, and that the computational cost to implement BSL does not vary drastically as the dimension of the summaries and sample size increases.

**Toad Example**

This example is an individual-based model of a species called Fowler’s Toads (*Anaxyrus fowleri*) developed by Marchand, Boenke, and Green (2017), which has also been previously analyzed by An, Nott, and Drovandi (2020). The example is briefly described here; see Marchand, Boenke, and Green (2017) and An, Nott, and Drovandi (2020) for further details.

The model assumes that a toad hides in its refuge site in the daytime and moves to a randomly chosen foraging place at night. GPS location data are collected on $n_t$ toads for $n_d$ days, so the matrix of observations $Y$ is $n_d \times n_t$ dimensional. This example uses both simulated and real data. The simulated data uses $n_t = 66$ and $n_d = 63$ and we summarize the data by four sets of statistics comprising the relative moving distances for time lags of 1, 2, 4, and 8 days. For instance, $y_1$ consists of the displacement information of lag 1 day, $y_1 = \{|Y_{ij} - Y_{i+1j}|; 1 \leq i \leq n_d - 1, 1 \leq j \leq n_t\}$.

Simulating from the model involves two processes. For each toad, we first generate an overnight displacement, $\Delta y$, then mimic the returning behavior with a simplified model. The overnight displacement is assumed to belong to the Lévy-alpha stable distribution family, with stability parameter $\alpha$ and scale parameter $\delta$. With probability $1 - p_0$, the toad takes refuge at the location it moved to. With probability $p_0$, the toad returns to the same refuge site as day $i$, $1 \leq i \leq M$ (where $M$ is the number of days the simulation has run for), where $i$ is selected randomly from $1, 2, \ldots, M$ with equal probability. For the simulated data, $\theta = (\alpha, \delta, p_0) = (1.7, 35, 0.6)$, which is a parameter value fitting the real data well, and assume a uniform prior over $(1, 2) \times (0, 100) \times (0, 0.9)$ for $\theta$.

As in Marchand, Boenke, and Green (2017), the dataset of displacements is split into two components. If the absolute value of the displacement is less than 10 meters, it is assumed the toad has returned to its starting location. For the summary statistic, we consider the number of toads that returned. For the non-returns (absolute displacement greater than 10 meters), we calculate the log difference between adjacent $p$-quantiles with $p = 0, 0.1, \ldots, 1$ and also the median. These statistics are computed separately for the four time lags, resulting in a 48-dimensional statistic. For standard BSL, $m = 500$ simulations are used per MCMC iteration. However, with a shrinkage parameter of $\gamma = 0.1$, it was only necessary to use $m = 50$ simulations per MCMC iteration. For the adjustment, $m = 500$ is used as far fewer synthetic likelihood evaluations are required. For the simulated data, the MCMC acceptance rates are 16% and 21% for standard and shrinkage BSL, respectively. For the real data, the acceptance rates are both roughly 24%.

**Figure 2** summarizes the results for the simulated data and shows that the shrinkage BSL posterior underestimates the variance and has the wrong dependence structure compared to the standard BSL posterior. The adjusted posterior produces uncertainty quantification that is closer to the standard BSL procedure, although its larger variances indicate that there is a loss in efficiency in using frequentist inference based on the shrinkage BSL point estimate. The results for the real data in **Figure 3** are qualitatively similar.

---

**Figure 2.** Adjustment results for the toad example based on the simulated data. The panels in the top row are bivariate contour plots of the standard and shrinkage BSL posteriors. The panels in the bottom row are bivariate contour plots of the standard and adjusted BSL posteriors.
In Section B.2 of the supplementary material, we further investigate the performance of BSL across repeated samples. The results demonstrate that point estimators obtained from standard BSL have superior performance, but that the credible sets result in some under-coverage, with shrinkage BSL producing significant under-coverage. In contrast, the adjusted BSL approach displays coverages that are closest to the nominal level.

6. Discussion

Our article, and earlier research, demonstrates empirically that BSL is a useful alternative to ABC. Consequently, it is important to study the theoretical behavior of BSL. Under reasonable assumptions, we demonstrate that the BSL posterior can correctly quantify uncertainty. More generally, our results show that the large sample behavior of the BSL posterior is similar to ABC posteriors that use algorithmic settings which yield correct uncertainty quantification. We also examine the effect of estimating the mean and covariance matrix in synthetic likelihood algorithms, and we theoretically demonstrate that BSL is more computationally efficient than ABC.

Adjustments are also discussed for a misspecified summary statistic covariance matrix in the synthetic likelihood. These adjustments may also be useful when the model for y is misspecified, and inference on the pseudo-true parameter is of interest. Our adjustment methods do not help correct inference in the case where the summary statistics are not normal. Some approaches consider more complex parametric models than the normal for addressing this issue, and the asymptotic framework developed here could be adapted to other parametric model approximations for the summaries. These extensions are left to future work.

Although our adjustments could be useful when the model for y is misspecified, it is helpful to distinguish different types of misspecification. Model incompatibility is said to occur when it is impossible to recover the observed summary statistic for any \( \theta \), but we do not investigate the behavior of synthetic likelihood in detail in this case. Frazier and Drovandi (2021) and Frazier, Robert, and Rousseau (2020) demonstrate that standard BSL and ABC can both perform poorly under incompatibility. Frazier and Drovandi (2021) propose some extensions to BSL allowing greater robustness and computational efficiency in this setting. More research is needed to compare BSL and ABC when model incompatibility occurs.

Supplementary Materials

The supplementary material contains proofs of all technical results presented in the article, additional numerical results, and further discussion on the differences between the BSL posterior and related classes of posterior distributions. All code needed to run the experiments given in the article is now available at https://github.com/cdrovandi/BSL-asymptotics.

Acknowledgments

We thank Ziwen An for preparing computer code for the toad example. We are grateful to the Associate Editor and two referees for helpful comments. All remaining errors are our own.

Funding

David Frazier was supported by the Australian Research Council's Discovery Early Career Researcher Award funding scheme (DE200101070). David Nott was supported by a Singapore Ministry of Education Academic
References

An, Z., Nott, D. J., and Drovandi, C. (2020), “Robust Bayesian Synthetic Likelihood via a Semi-parametric Approach,” *Statistics and Computing*, 30, 543–557. [2822,2830]

An, Z., South, L. F., Drovandi, C. C., and Nott, D. J. (2019), “Accelerating Bayesian Synthetic Likelihood with the Graphical Lasso,” *Journal of Computational and Graphical Statistics*, 28, 471–475. [2822]

Andrieu, C., and Roberts, G. O. (2009), “The Pseudo-Marginal Approach for Efficient Monte Carlo Computations,” *The Annals of Statistics*, 37, 697–725. [2823]

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002), “Approximate Bayesian Computation in Population Genetics,” *Genetics*, 162, 2025–2035. [2829]

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016), “A General Framework for Updating Belief Distributions,” *Journal of the Royal Statistical Society*, Series B, 78, 1103–1130. [2822,2823]

Chaudhuri, S., Ghosh, S., Nott, D. J., and Pham, K. C. (2020), “On a Variational Approximation Based Empirical Likelihood ABC Method,” arXiv:2011.07721. [2822]

Chernozhukov, V., and Hong, H. (2003), “An MCMC Approach to Classical Estimation,” *Journal of Econometrics*, 115, 293–346. [2822,2823]

Deligiannidis, G., Doucet, A., and Pitt, M. K. (2018), “The Correlated Pseudo-Marginal Method,” *Journal of the Royal Statistical Society*, Series B, 80, 839–870. [2823]

Doucet, A., Pitt, M. K., Deligiannidis, G., and Kohn, R. (2015), “Efficient Implementation of Markov chain Monte Carlo When Using an Unbiased Likelihood Estimator,” *Biometrika*, 102, 295–313. [2823]

Drovandi, C. C., Pettitt, A. N., and Lee, A. (2015), “Bayesian Indirect Inference Using a Parametric Auxiliary Model,” *Statistical Science*, 30, 72–95. [2822]

Everitt, R. G. (2017), “Boosted Synthetic Likelihood,” arXiv:1711.05825. [2822]

Fasiolo, M., Wood, S. N., Hartig, F., and Bravington, M. V. (2018), “An Extended Empirical Saddlepoint Approximation for Intractable Likelihoods,” *Electronic Journal of Statistics*, 12, 1544–1578. [2822]

Fornier, J.-J., and Ng, S. (2018), “The ABC of Simulation Estimation with Auxiliary Statistics,” *Journal of Econometrics*, 205, 112–139. [2822]

Frazier, D. T., and Drovandi, C. (2021), “Robust Approximate Bayesian Inference with Synthetic Likelihood,” *Journal of Computational and Graphical Statistics*, 30, 958–976. [2831]

Frazier, D. T., Martin, G. M., Robert, C. P., and Rousseau, J. (2018), “Asymptotic Properties of Approximate Bayesian Computation,” *Biometrika*, 105, 593–607. [2821,2824,2826]

Frazier, D. T., Robert, C. P., and Rousseau, J. (2020), “Model Misspecification in Approximate Bayesian Computation: Consequences and Diagnostics,” *Journal of the Royal Statistical Society*, Series B, 82, 421–444. [2822,2831]

Gutmann, M. U., and Corander, J. (2016), “Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models,” *Journal of Machine Learning Research*, 17, 1–47. [2822]

Hannan, E. J. (1976), “The Asymptotic Distribution of Serial Covariances,” *The Annals of Statistics*, 4, 396–399. [2829]

Kong, A. (1992), “A Note on Importance Sampling Using Standardized Weights,” Chicago Department of Statistics Technical Report 348. [2829]

Kreiss, J.-P., and Paparoditis, E. (2011), “Bootstrap Methods for Dependent Data: A Review,” *Journal of the Korean Statistical Society*, 40, 357–378. [2827]

Li, W., and Fearnhead, P. (2018a), “Convergence of Regression-Adjusted Approximate Bayesian Computation,” *Biometrika*, 105, 301–318. [2821,2824,2825,2826]

Li, W., and Fearnhead, P. (2018b), “On the Asymptotic Efficiency of Approximate Bayesian Computation Estimators,” *Biometrika*, 105, 285–299. [2821,2825]

Marchand, P., Boenke, M., and Green, D. M. (2017), “A Stochastic Movement Model Re reproduces Patterns of Site Fidelity and Long-Distance Dispersal in a Population of Fowler’s Toads (*Anaxayrus fowleri)*,” *Ecological Modelling*, 360, 63–69. [2830]

McKay, M., Beckman, R., and Conover, W. (1979), “Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code,” *Technometrics*, 21, 239–245. [2828]

Meeds, E., and Welling, M. (2014), “GPS-ABC: Gaussian Process Surrogate Approximate Bayesian Computation,” in *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, UAI’14, Arlington, VA, pp. 593–602. AUAI Press. [2822]

Mengersen, K. L., Pudlo, P., and Robert, C. P. (2013), “Bayesian Computation via Empirical Likelihood,” *Proceedings of the National Academy of Sciences*, 110, 1321–1326. [2822]

Müller, U. K. (2013), “Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix,” *Econometrica*, 81, 1805–1849. [2822]

Ong, V. M.-H., Nott, D. J., Tran, M.-N., Sisson, S., and Drovandi, C. (2018a), “Variational Bayes with Synthetic Likelihood,” *Statistics and Computing*, 28, 971–988. [2822]

Pitt, M. K., Silva, R. d. S., Giordani, P., and Kohn, R. (2012), “On Some Properties of Markov chain Monte Carlo Simulation Methods Based on the Particle Filter,” *Journal of Econometrics*, 171, 134–151. [2821]

Price, L. F., Drovandi, C. C., Lee, A. C., and Nott, D. J. (2018), “Bayesian Synthetic Likelihood,” *Journal of Computational and Graphical Statistics*, 27, 1–11. [2821,2822,2823,2826]

Priddle, J. W., Sisson, S. A., Frazier, D. T., Turner, I., and Drovandi, C. (2022), “Efficient Bayesian Synthetic Likelihood with Whitening Transformations,” *Journal of Computational and Graphical Statistics*, 31, 50–63. [2821,2822]

Sherlock, C., Thiery, A. H., Roberts, G. O., and Rosenthal, J. S. (2015), “On the Efficiency of Pseudo-Marginal Random Walk Metropolis Algorithms,” *The Annals of Statistics*, 43, 238–257. [2823]

Sisson, S. A., Fan, Y., and Beaumont, M. (2018), *Handbook of Approximate Bayesian Computation*, Boca Raton, FL: Chapman and Hall/CRC. [2821]

Thomas, O., Dutta, R., Corander, J., Kaski, S., and Gutmann, M. U. (2022), “Likelihood-Free Inference by Ratio Estimation,” *Bayesian Analysis*, 17, 1–31. [2822]

Tran, M.-N., Kohn, R., Qiuroz, M., and Villani, M. (2016), “The Block Pseudo-Marginal Sampler,” arXiv preprint arXiv:1603.02485. [2823]

Vershynin, R. (2018), *High-Dimensional Probability: An Introduction with Applications in Data Science* (Vol. 47), Cambridge: Cambridge University Press. [2824]

Warton, D. I. (2008), “Penalized Normal Likelihood and Ridge Regularization of Correlation and Covariance Matrices,” *Journal of the American Statistical Association*, 103, 340–349. [2828]

Wilkinson, R. (2014), “Accelerating ABC Methods Using Gaussian Processes,” *Journal of Machine Learning Research*, 33, 1015–1023. [2822]

Wood, S. N. (2010), “Statistical Inference for Noisy Nonlinear Ecological Dynamic Systems,” *Nature*, 466, 1102–1107. [2822]