Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Incremental Mixture Importance Sampling With Shotgun Optimization

Biljana Jonoska Stojkova & David A. Campbell

To cite this article: Biljana Jonoska Stojkova & David A. Campbell (2019) Incremental Mixture Importance Sampling With Shotgun Optimization, Journal of Computational and Graphical Statistics, 28:4, 806-820, DOI: 10.1080/10618600.2019.1592756

To link to this article: https://doi.org/10.1080/10618600.2019.1592756

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2019, VOL. 28, NO. 4, 806–820
https://doi.org/10.1080/10618600.2019.1592756

Taylor & Francis
Taylor & Francis Group

Incremental Mixture Importance Sampling With Shotgun Optimization

Biljana Jonoska Stojkova[a] and David A. Campbell[b]

[a]Department of Statistics, University of British Columbia, Vancouver, Canada; [b]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, Canada

ABSTRACT

This article proposes a general optimization strategy, which combines results from different optimization or parameter estimation methods to overcome shortcomings of a single method. Shotgun optimization is developed as a framework which employs different optimization strategies, criteria, or conditional targets to enable wider likelihood exploration. The introduced shotgun optimization approach is embedded into an incremental mixture importance sampling algorithm to produce improved posterior samples for multimodal densities and creates robustness in cases where the likelihood and prior are in disagreement. Despite using different optimization approaches, the samples are combined into samples from a single target posterior. The diversity of the framework is demonstrated on parameter estimation from differential equation models employing diverse strategies including numerical solutions and approximations thereof. Additionally the approach is demonstrated on mixtures of discrete and continuous parameters and is shown to ease estimation from synthetic likelihood models. R code of the implemented examples can be found at *https://github.com/BiljanaJSJ/IMIS-ShOpt*. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2017
Revised December 2019

KEYWORDS
Chaotic stochastic difference models; Differential equation models; Importance sampling; Multimodal posterior; Numerical optimization; Synthetic likelihood topologies

1. Introduction

Sampling from a posterior density is challenging when the posterior modes are separated with deep valleys of low probability or when the posterior space is rife with many minor modes, ripples, and ridges. Theoretically, standard Metropolis–Hastings or Gibbs algorithms converge to the target density if run infinitely long. Tempering methods, such as simulated tempering (Marinari and Parisi 1992; Geyer and Thompson 1995; Zhang and Ma 2008) and parallel tempering (Swendsen and Wang 1986; Geyer 1991; Hukushima and Nemoto 1996), are random-walk variants designed to efficiently deal with sampling from multimodal distributions. However, parallel tempering could exacerbate topological challenges of the posterior if the prior is inconsistent with the likelihood, trapping the sampler in a local mode (Campbell and Steele 2012).

Importance sampling algorithms, such as sampling importance resampling (SIR) (Rubin 1987, 1988; Poole and Raftery 2000; Alkema et al. 2011) or sequential Monte Carlo variants (SMC) (Del Moral, Doucet, and Jasra 2006) take advantage of computing the sampling weights in parallel. The difficulty with importance sampling methods is choosing the initial importance density to cover the important modes of the target density. The prior is often chosen to be this initial importance density, but in practice the challenge of obtaining an initial sample in the relevant region of the posterior is difficult unless the prior is close to the posterior.

A frequentist alternative to MCMC methods would be to use optimization but in the presence of multiple isolated modes, different starting points for the optimizer result in multiple optima. Then the problem shifts to finding a way to combine these local optima.

Incremental mixture importance sampling with optimization (IMIS-Opt) (Raftery and Bao 2010) is designed to discover important posterior modes by using the prior as a starting point for optimization, and then building a posterior by incrementally adding optimized local posterior approximations. Sums of Gaussians can be used to approximate any continuous probability distribution in L^1 and L^∞ norms simultaneously (Bacharoglou 2010) leaving the problem finding the appropriate Gaussians and the weights for the mixture. The optimization stage of IMIS-Opt and its use of local geometry improve the exploration and approximation of the posterior. However, if the prior disagrees with the likelihood, that is, if the prior covers the basin of attraction of local but not global likelihood modes, then IMIS-Opt will miss important modes. As a remedy, one can choose a diffuse prior, but this implies that the prior should be chosen for algorithmic convenience rather than representing expert opinion.

In this article, we modify the IMIS-Opt algorithm by replacing the optimization step with a general optimization strategy, which is based on the no free lunch (NFL) theorem for optimization (Wolpert and Macready 1997); no single optimization method outperforms other methods in every problem but some problems have special structure which, if known can offer improvements when exploited by the optimization algorithm.

The proposed multiple-method optimization strategy balances discovery of global and local modes by combining results from multiple parameter estimation methods, which may arise

CONTACT Biljana Jonoska Stojkova b.stojkova@stat.ubc.ca Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JCGS.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

from exploiting different types of problem structure or from using different black box algorithms. We refer to this strategy as shotgun optimization (ShOpt), and the resulting algorithm as Incremental mixture importance sampling with shotgun optimization (IMIS-ShOpt).

The way the ShOpt combines results from different competing methods is substantially different from multi-objective optimization (Kuhn and Tucker 1951; Miettinen 2012). While multi-objective optimization is designed to optimize simultaneously several objectives, the ShOpt strategy is a single objective optimization targeted indirectly through independently optimizing over multiple criteria. For example, in inference from differential equation models, gradient optimization of likelihoods centered on numerical solutions to the differential equation typically end up in local, biased modes. However, likelihoods constructed through smoothing based strategies typically have wider variance parameter estimates with lower bias. While different optimization strategies provide different results, combined these approaches will explore widely for global modes while still exploring smaller, local modes in case they are important (Berger, Liseo, and Wolpert 1999; Walley and Moral 1999). All results are combined to build an importance distribution which is then compared to a single target posterior.

Shotgun optimization is analogous to the ensemble methods (Madigan and Raftery 1994; Hoeting et al. 1999; Friedman, Hastie, and Tibshirani 2001; Mendes-Moreira et al. 2012; Montgomery, Hollenbach, and Ward 2012) where relative importance of the predictions are determined using a combination of models. Ensemble methods rely on the notion that no particular model can fully capture the data features. Hence, some models better predict certain features of the data, while producing biased predictions in some areas. The ensemble methods overcome this induced bias by combining models together. In the ShOpt, certain methods provide better access to different posterior modes, and combining results from different methods overcomes the problem of introduced bias.

The rest of the article is organized as follows. Section 2 gives an overview of IMIS-Opt. In Section 3, ShOpt is introduced and IMIS-ShOpt is presented. Section 4 shows how to tailor the IMIS-ShOpt to a variety of models, each exploiting different model or parameter types. Section 4.1.1 is an easy to visualize one-dimensional FitzHugh–Nagumo example where the prior is inconsistent with the likelihood. Section 4.1.2 demonstrates IMIS-ShOpt overcomes complex posterior topology on the full FitzHugh–Nagumo differential equation model. Section 4.2 has an example with mixture of discrete and continuous parameters. The example in Section 4.3 uses a synthetic likelihood with a stochastic optimization strategy. Section 5 follows with discussion.

2. Incremental Mixture Importance Sampling With Optimization

The main objective of IMIS-Opt (Raftery and Bao 2010) is to iteratively construct an importance sampling distribution. The initial stage of the IMIS-Opt starts by drawing N_0 samples $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_0}\}$ from the prior and then calculating their weights based on the likelihood function. In the optimization stage, the D highest-weight points are selected to sequentially initialize the optimizer, which searches for the nearest mode in the target posterior space. Then B points, drawn from the multivariate Gaussian distribution centered at the modes found by the optimizer, are added to the current importance distribution. Weighting and sampling steps of the importance stage are iterated until the importance weights are reasonably uniform. After the stopping criterion is met, J inputs are resampled with replacement from $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_K}\}$ with weights $[w_1, \ldots, w_{N_K}]$ where N_K is the total number of particles after the kth stage of the importance sampling distribution. The pseudo-code of the IMIS-Opt is given in Algorithm 1.

If optimization and importance sampling stages are excluded, then the algorithm becomes a SIR algorithm (Rubin 1987, 1988; Poole and Raftery 2000; Alkema, Raftery, and Clark 2007). By excluding the optimization step, the algorithm becomes IMIS (Hesterberg 1995; Steele, Raftery, and Emond 2006).

IMIS-Opt initializes the optimizer using the D highest-weight points which allows it to include additional samples from potentially important regions of the parameter space. However, the successful mixing of the IMIS-Opt depends heavily on the consistency of the information in the prior and likelihood, and consequently, on whether or not samples from the prior are contained within basins of attraction for all the important posterior modes. The implication is that arguably, the prior should be chosen for optimization convenience rather than summarizing expert knowledge.

3. Incremental Mixture Importance Sampling With Shotgun Optimization

The success of IMIS-Opt depends heavily on the consistency between the prior and the likelihood. If the prior is inconsistent with the likelihood, all D optimization steps may be initialized in the basin of attraction of a local mode. The IMIS-ShOpt builds on IMIS-Opt, by altering the optimization stage to incorporate the ShOpt strategy, which consists of Q different competitive parameter estimation methods or optimization strategies. This implies using a variety of optimization methods or a fixed optimizer on variants of the function to optimize. The ShOpt strategy initializes Q different optimization methods (which could be run in parallel) for each of the D maximum weight points from the prior. Replacing the optimization step in Algorithm 1 with the ShOpt in Algorithm 2 gives the pseudo code of IMIS-ShOpt.

IMIS-ShOpt explores modes and merges the samples from different regions of the target posterior as explored by the variety of criteria. The modification of the optimization strategy depends on the parameter estimation method used. For example, if a parameter of interest is a location parameter, the multiple-method optimization in IMIS-ShOpt could be comprised of $Q = 2$ strategies: maximum likelihood method and method of moments. The posterior modifications targeted by different optimization strategies may give different results due to their inherent differences in topology of the posterior space. Weights in the resampling stage of IMIS-ShOpt ensure that points are appropriately sampled in the final stage. Hence, keeping the unlikely points in the importance sampling

```
Algorithm 1 IMIS-Opt
Goal: Draw samples from the target distribution $P(\boldsymbol{\theta} \mid \boldsymbol{Y})$.
Input: Data, model, likelihood function, prior distribution, $B$—the number of incremental points, $D$—the number of different initial points for the optimization, $N_0$—the number of the initial samples from the prior and $J$—the number of resampled points.
Initial stage: Draw $N_0$ samples $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \ldots, \boldsymbol{\theta_{N_0}}\}$ from the prior distribution $P(\boldsymbol{\theta})$, set $k = 0$.
  For each $\{\boldsymbol{\theta_i}, i = 1, \ldots, N_0\}$ calculate the sampling weights:
      $w_i^* = \frac{P(\boldsymbol{Y} \mid \boldsymbol{\theta_i})}{\sum_{j=1}^{N_0} P(\boldsymbol{Y} \mid \boldsymbol{\theta_j})}$   (1)
  Optimization stage:
  for $d = 1 : D$ do
    Use $\boldsymbol{\theta}^{(\text{initial})} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, \boldsymbol{w}^*(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{d-1}$ to initialize the optimizer and get local posterior maxima $\boldsymbol{\theta}_d^{(\text{Opt})} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, P(\boldsymbol{\theta} \mid \boldsymbol{Y})$ along with the corresponding inverse negative Hessian $\boldsymbol{\Sigma}_d^{(\text{Opt})}$.
    Construct $\boldsymbol{\Theta}_d$ from $\boldsymbol{\Theta}_{d-1}$ by excluding $\frac{N_0}{D}$ nearest neighbor points, that is, exclude $\frac{N_0}{D}$ points $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{d-1}$ that minimize the Mahalanobis distance,
      $(\boldsymbol{\theta} - \boldsymbol{\theta}_d^{(\text{Opt})})'(\boldsymbol{\Sigma}_d^{(\text{Opt})})^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_d^{(\text{Opt})}).$   (2)
    Draw B samples $\boldsymbol{\theta}_{1:B} \sim \text{MVN}(\boldsymbol{\theta}_{\text{d}}^{(\text{Opt})}, \boldsymbol{\Sigma}_{\text{d}}^{(\text{Opt})})$; add these samples to the importance sampling distribution and evaluate $H_k = \text{MVN}(\boldsymbol{\theta}_{1:\text{B}} \mid \boldsymbol{\theta}_{\text{d}}^{(\text{Opt})}, \boldsymbol{\Sigma}_{\text{d}}^{(\text{Opt})})$.
  end for
  Importance sampling stage:
  For each $\{\boldsymbol{\theta_i}, i = 1, \ldots, N_k\}$ calculate weights,
      $w_i^{(k)} = \frac{cP(\boldsymbol{Y} \mid \boldsymbol{\theta_i})P(\boldsymbol{\theta_i})}{\frac{N_0}{N_k}P(\boldsymbol{\theta_i}) + \frac{B}{N_k}\sum_{s=1}^{k} H_s(\boldsymbol{\theta_i})},$   (3)
  where $N_k = N_0 + B(D + k)$ and $c = 1/\sum_{i=1}^{N_k} w_i^{(k)}$ is the normalizing constant.
  while $\sum_{1}^{N_k}(1 - (1 - w^{(k)})^J) < J(1 - \exp(-1))$, that is, importance sampling weights are not approximately uniform do
    Update $k = k + 1$.
Algorithm 1 *
Algorithm 1 IMIS-Opt—continued
    Choose the maximum weight input $\boldsymbol{\theta}_k$ and estimate $\boldsymbol{\Sigma}_k$ as the weighted covariance of B inputs with smallest Mahalanobis distance,
      $w_p^{-1}(\boldsymbol{\theta})\,(\boldsymbol{\theta} - \boldsymbol{\theta}_k)'\,(\boldsymbol{\Sigma_\pi})^{-1}\,(\boldsymbol{\theta} - \boldsymbol{\theta}_k),$
    where the weights $w_p(\boldsymbol{\theta})$ are proportional to the average of the importance weights and the uniform weights $\frac{1}{N_k}$, $\boldsymbol{\Sigma}_\pi$ is the covariance of the initial importance distribution.
    Draw B samples $\boldsymbol{\theta}_{1:B} \sim \text{MVN}(\boldsymbol{\theta}_{\text{k}}, \boldsymbol{\Sigma}_{\text{k}})$; add these points to the importance sampling distribution and evaluate $H_k = \text{MVN}(\boldsymbol{\theta}_{1:\text{B}} \mid \boldsymbol{\theta}_{\text{k}}, \boldsymbol{\Sigma}_{\text{k}})$.
    Update weights $w^{(k)}$ using Equation (3).
  end while
  Resampling stage:
  Resample J points with replacement from $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_k}\}$ and weights $(w_1, \ldots, w_{N_k})'$.
```

distribution does not harm the algorithm, but it does improve the posterior exploration.

IMIS-ShOpt has several user defined control parameters, the initial sample size N_0, the number of starting points for the optimizer D, the number of optimization strategies Q, the number of points added with every local multivariate normal approximation B, and the final number of resampled points, J. Being an importance sampling algorithm, IMIS-ShOpt is unbiased for any choice of control parameters (Raftery and Bao 2010), however, the choice of control parameters can affect its efficiency. Raftery and Bao (2010) found that the choice of $N_0 = 1000d$, $B = 100d$, and $J = 3000$, where d is the number of parameters estimated, gives good results for estimating marginal likelihoods. For posterior inference, we have found similar success with IMIS-ShOpt, but tend to use $J \geq 10{,}000$ so as to reduce Monte Carlo variation in tails and interval estimates. The choice of Q is problem specific and typically has a natural value that depends on the class of models (see Section 4). In the absence of creative problem specific strategies, we have found practical success with using $Q = 3$ different optimizers targeting a single criteria. Increasing N_0, D, B, and Q increases the number of samples and will, therefore, improve the importance distribution.

IMIS-ShOpt builds a mixture importance distribution incrementally from a mixture of the prior and a set of Gaussians in under-represented locations at the current stage of the algorithm. As such the algorithm behaves by effectively monitoring its own convergence. As with other IMIS algorithms, poor coverage of the target posterior can be detected by presence of large weights. The algorithm stopping condition is when the importance sample is built up to J uniformly weighted particles, or equivalently, when the expected number of unique points after the final resampling of J points is at least $J(1 - e^{-1})$.

Following (Raftery and Bao 2010), we assess the performance of IMIS-ShOpt by monitoring the following criteria:

- the maximum importance weight among the N_k inputs, when converged this is near $1/N_k$;
- the variance of the rescaled importance weights in units of N_k^{-2}, $\hat{V}(w) = \frac{1}{N_k}\sum_{i=1}^{N_k}(N_k w_i - 1)^2$, when converged this is close to 0;
- the entropy of the importance weights relative to uniformity $\hat{U}(w) = -\sum_{i=1}^{N_k} w_i \frac{\log(w_i)}{\log(N_k)}$, when converged this is close to 1.
- the effective sample size $\text{ESS(w)} = \frac{1}{\sum_1^{\text{N}_\text{k}} \text{w}_\text{i}^2}$, when converged this is close to N_k.
- the expected number of unique weights $\hat{Q} = \sum_{i=1}^{N_k}(1 - (1 - w_i)^J)$, the stopping criterion is when $\hat{Q} \geq J(1 - e^{-1})$.

Algorithm 2 The Shotgun optimization

```
Optimization stage:
for $d = 1 : D$ do
    Find the $d$th maximum weight point $\boldsymbol{\theta}_d^{(\text{initial})} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, \boldsymbol{w}^{(k)}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}_{d-1}$ to initialize $Q$ optimizers.
    for $q = 1 : Q$ do
        Use q-th optimization method initialized at $\boldsymbol{\theta}_d^{(\text{initial})}$ to obtain local maxima $\boldsymbol{\theta}_{d,q}^{(\text{Opt})}$ along with the corresponding inverse negative Hessian $\boldsymbol{\Sigma}_{d,q}^{(\text{Opt})}$ (this step can be parallelized).
        Construct $\boldsymbol{\Theta}_d$ from $\boldsymbol{\Theta}_{d-1}$ by excluding $\frac{N_0}{QD}$ nearest neighbor points, that is, exclude $\frac{N_0}{QD}$ points $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_{d-1}$ that minimize the Mahalanobis distance,
            $$(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{d,q}^{(\text{Opt})})'(\boldsymbol{\Sigma}_{d,q}^{(\text{Opt})})^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{d,q}^{(\text{Opt})}). \quad (4)$$
        Draw B samples $\boldsymbol{\theta}_{1:B} \sim \text{MVN}(\boldsymbol{\theta}_{\text{d,q}}^{(\text{Opt})}, \boldsymbol{\Sigma}_{\text{d,q}}^{(\text{Opt})})$; add these points to the importance sampling distribution and evaluate $H_k = \text{MVN}(\boldsymbol{\theta}_{1:\text{B}} \mid \boldsymbol{\theta}_{\text{d,q}}^{(\text{Opt})}, \boldsymbol{\Sigma}_{\text{d,q}}^{(\text{Opt})})$.
    end for
end for
```

To avoid numerical instabilities, the weights given in Equation (3) are calculated as $w_i = \frac{\exp(\tilde{w}_i)}{\sum_{j=1}^{N_k}\exp(\tilde{w}_j)}$, where $\tilde{w}_i = \log(P(\boldsymbol{Y} \mid \boldsymbol{\theta})) + \log P(\boldsymbol{\theta}) - \log\{\frac{N_0}{N_k}P(\boldsymbol{\theta_i}) + \frac{B}{N_k}\sum_{s=1}^{k} H_s(\boldsymbol{\theta_i})\} - \log(C)$, and C is a constant obtained as a maximum of the $P(\boldsymbol{Y} \mid \boldsymbol{\theta_j})$.

IMIS-ShOpt is a form of defensive sampling (Hesterberg 1995) which uses a mixture density to avoid having an importance distribution whose density is too low over portions of the target. A mixture also defends against infinite variance marginal likelihood estimators by increasing the variance of the importance density (Owen and Zhou 2000). As an integrated likelihood estimator for model selection purposes, IMIS-ShOpt is unbiased, strongly consistent, and asymptotically normal if B grows at the Kth stage toward infinity (Raftery and Bao 2010). The ShOpt strategy does not degrade the assumptions behind their analysis. Although finite sample sizes cannot rule out the possibility of missed modes, the practical setting with constant B provides a means of widely exploring $P(\theta \mid \mathbf{Y})$ because ShOpt improves the chances of finding distant modes.

IMIS-ShOpt is related to adaptive multiple importance sampling (AMIS) in that weights are updated at each iteration of the algorithm, but differs in that AMIS does not include an optimization stage and AMIS uses all of the N_k particles at stage k to determine the next H_k, whereas IMIS-ShOpt uses the nearest B samples to the highest weighted point in its construction of H_k. Because of the sequential dependence structure imposed by adding importance samples in regions found to be underrepresented, a more general theoretical convergence for AMIS and IMIS-ShOpt algorithms to the target posterior remains an open problem (Cornuet et al. 2012; Marin, Pudlo, and Sedki 2012; Sbert, Havran, and Szirmay-Kalos 2018). However, the special case where the incremental distributions are determined a priori convergences to the target posterior under mild conditions (Douc et al. 2007).

4. Examples

In a simple interpretation of the ShOpt strategy one could use, for example, conjugate gradient, Nelder–Mead simplex, and Newton's method as the $Q = 3$ optimization approaches targeting a single optimization criterion. Instead, in the following examples we customize IMIS-ShOpt to the nuances of the models by targeting Q different optimization criteria. We apply the IMIS-ShOpt to three models highlighting the ability to overcome common challenges in statistics. The FitzHugh–Nagumo example is an ordinary differential equation (ODE), of the form $\frac{dX(t)}{dt} = f(X(t))$, where our likelihood is multimodal. There are several established methods for estimating parameters from differential equation models that we exploit within IMIS-ShOpt. Using an easy to visualize one-dimensional variant of the FitzHugh–Nagumo model, Section 4.1.1 showcases the IMIS-ShOpts ability to overcome the problem of a prior centred on a distant minor likelihood mode. The full FitzHugh–Nagumo model is considered in Section 4.1.2 showcasing more general applicability of the ShOpt approach in differential equation models. A susceptible-infectious-recovered epidemiological example is used in Section 4.2 with a mixture of continuous and discrete parameters. Finally, Section 4.3 uses a synthetic-likelihood example with a stochastic choice of optimization criterion.

4.1. FitzHugh-Nagumo Model

The FitzHugh–Nagumo (FhN) model (FitzHugh 1961; Nagumo, Arimoto, and Yoshizawa 1962) captures the behavior of spike potentials in the giant axon of squid neurons. The FhN model is described by a system of two nonlinear ordinary differential equations corresponding to the voltage across the membrane, $V(t)$, and outward currents (recovery), $R(t)$, at time t with a vector of parameters of interest $\boldsymbol{\theta} = [a, b, c]$,

$$\begin{aligned} \frac{dV}{dt} &= c\left(V(t) - V(t)^3/3 + R(t)\right) \quad \text{and} \\ \frac{dR}{dt} &= -\frac{1}{c}\left(V(t) - a + bR(t)\right). \end{aligned} \tag{5}$$

An analytic solution of the ODE system in Equation (5) does not exist but a numerical solution can be produced with initial states values $V(0)$ and $R(0)$ which must be appended to the parameter vector $\boldsymbol{\theta}$. In the following two examples, the measurement error model for data $\boldsymbol{Y}_V(t)$ and $\boldsymbol{Y}_R(t)$ is centered about $V(\boldsymbol{\theta}, t)$ and $R(\boldsymbol{\theta}, t)$, the numerical solution of Equation (5),

$$\begin{aligned} \boldsymbol{Y}_V(t) \mid \boldsymbol{\theta} &\sim N\left(V(\boldsymbol{\theta}, t), \sigma_V^2\right) \quad \text{and} \\ \boldsymbol{Y}_R(t) \mid \boldsymbol{\theta} &\sim N\left(R(\boldsymbol{\theta}, t), \sigma_R^2\right). \end{aligned} \tag{6}$$

The ShOpt strategy used combines three different parameter estimation strategies tailored to ODE models: (i) Nonlinear least squares (NLS) (Bates and Watts 1988; Seber and Wild 1989), (ii) two stage estimator (Varah 1982; Brunel 2008; Liang and Wu 2008) and (iii) generalized profiling (GP) (Ramsay et al. 2007). All three are described bellow. The results from these $Q = 3$

optimization strategies are combined and compared using a single target posterior based on the likelihood in Equation (6).

Nonlinear Least Squares. Following Bates and Watts (1988), the maximum likelihood estimator for $\hat{\boldsymbol{\theta}}$ is obtained by minimizing the negative log-likelihood for observations y_{sj} over system states $s \in \{V, R\}$, where using our Gaussian likelihood results in minimizing the sum of squared residuals:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{s=1}^{S} \sum_{j=1}^{n_s} \left[y_{sj} - \boldsymbol{X}(\boldsymbol{\theta}, t_{sj})\right]^2. \tag{7}$$

The NLS method has several drawbacks. In order to minimize Equation (7), NLS requires numerically solving Equation (5) at each evaluation of the likelihood and, therefore, requires including initial states as parameters. Results of NLS depend on the optimization initialization especially when Equation (7) exhibits multiple modes, as is common with ODE models (Campbell and Steele 2012).

Two-Stage Method. The two-stage method first smooths the data to estimate $\hat{\boldsymbol{X}}(\boldsymbol{\theta}, t)$ and its derivative $\widehat{\frac{d\boldsymbol{X}(\boldsymbol{\theta},t)}{dt}}$ (Varah 1982; Brunel 2008; Liang and Wu 2008). In the second stage, parameter estimates are obtained by maximizing fidelity of the smooth to the ODE model dynamics in Equation (6).

The local polynomial procedure (Fan and Gijbels 1996) approximates the sth state $\boldsymbol{X}_s(\boldsymbol{\theta}, t_{sj})$ by a νth order polynomial, in a neighborhood of the time point t_{s0}, with $a_i(\boldsymbol{\theta}, t_{s0}) = \boldsymbol{X}_s^{(i)}(\boldsymbol{\theta}, t_{s0})$ for $i = 0, \ldots, \nu$,

$$\begin{aligned}\boldsymbol{X}_s(\boldsymbol{\theta}, t_{sj}) &\approx \boldsymbol{X}_s(\boldsymbol{\theta}, t_{s0}) + (t_{sj} - t_{s0})\boldsymbol{X}_s^{(1)}(\boldsymbol{\theta}, t_{s0}) \\ &\quad + \cdots + (t_{sj} - t_{s0})^s \boldsymbol{X}_s^{(\nu)}(\boldsymbol{\theta}, t_{s0})/\nu! \\ &= \sum_{i=0}^{\nu} a_i(\boldsymbol{\theta}, t_{s0})(t_{sj} - t_{s0})^i, \\ &\qquad \text{for } s = 1, \ldots, S, j = 1, \ldots, n_s. \end{aligned} \tag{8}$$

Following Fan and Gijbels (1996), the estimators $\widehat{\boldsymbol{X}_s^{(i)}}(\boldsymbol{\theta}, t)$, $i = 0, 1$, are obtained by minimizing the locally weighted least-square criterion,

$$\sum_{j=1}^{n_s} \left[y_{sj} - \sum_{i=0}^{\nu} a_i (t_{sj} - t_{s0})^i \right]^2 K_h(t_{sj} - t_{s0}), \tag{9}$$

where h controls the size of the neighborhood around t_{s0}, $K_h(.) = K_h/h$ controls the weights, and $K(.)$ is a Kernel weight function.

In the second stage, $\hat{\boldsymbol{\theta}}$ is obtained by minimizing the sum of squared residuals between the derivative estimate and the ODE model,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{s=1}^{S} \sum_{j=1}^{n_s} \left[\frac{\widehat{d\boldsymbol{X}_s(\boldsymbol{\theta}, t_{sj})}}{dt} - f_s(\hat{\boldsymbol{X}}(\boldsymbol{\theta}, t_{sj}), \boldsymbol{\theta}) \right]^2. \tag{10}$$

While model (6) only allows noise at the state, here noise enters the system at both the state and derivative levels. The two-stage method is computationally more efficient than the NLS, since it avoids numerical solutions at each evaluation of the objective function. However, this gain of computational efficiency comes at the cost of accuracy. Namely, in the first stage the data are smoothed without using the ODE model information. The ODE model is only used in the second stage to obtain $\hat{\boldsymbol{\theta}}$ based on the first stage smoothing results. Separating the estimation procedure in two stages results in a reduced estimation accuracy of the ODE parameters (Ding and Wu 2014).

Generalized Profiling. Avoiding the numerical solution to the ODE system, the generalized profiling (GP) method uses model based data smoothing by penalizing deviation at the level of the derivative. The data smooth $\hat{X}(\theta) = \boldsymbol{\Phi}(t)\boldsymbol{C}$ is a basis expansion with coefficients $\boldsymbol{C} = [\boldsymbol{C}_1, \ldots, \boldsymbol{C}_s]$ and basis functions $\boldsymbol{\Phi}(t) = [\boldsymbol{\Phi}_1(t), \ldots, \boldsymbol{\Phi}_s(t)]$.

GP is a parameter cascade procedure which estimates θ through the profile likelihood, profiling over $\boldsymbol{C}$. At each evaluation of the profile likelihood, the basis coefficients are estimated holding θ fixed

$$\begin{aligned}\hat{\boldsymbol{C}} \mid \boldsymbol{\theta}, \lambda, \boldsymbol{Y} &= \arg\min_{\boldsymbol{C}} \sum_{s=1}^{S} \sum_{j=1}^{n_s} \left[y_{sj} - \boldsymbol{\Phi}_s(t_{sj})\boldsymbol{c}_{\boldsymbol{s}}\right]^2 \\ &\quad + \sum_{s=1}^{S} \lambda \int_{\boldsymbol{T}} \left[\frac{d\boldsymbol{\Phi}_s(t)\boldsymbol{c}_{\boldsymbol{s}}}{dt} - f_s(\boldsymbol{\Phi}(t)\boldsymbol{C}, \boldsymbol{\theta}) \right]^2 dt, \end{aligned} \tag{11}$$

where t is integrated over the interval of observation times. The first term of Equation (11) represents a sum of squares between the observed states and the basis expansion, while the second term measures the fidelity of the basis expansion to the ODE model. The smoothing parameter λ controls the trade-off between the two. The profile likelihood optimization uses an outer optimization criterion to estimate θ via

$$\hat{\boldsymbol{\theta}} \mid \boldsymbol{C}, \boldsymbol{Y} = \arg\min_{\boldsymbol{\theta}} \sum_{s=1}^{S} \sum_{j=1}^{n_s} \left[y_{sj} - \boldsymbol{\Phi}_s(t_{sj})\boldsymbol{c}_s(\boldsymbol{\theta})\right]^2. \tag{12}$$

Model based smoothing in Equation (11) makes GP a sort of hybrid between NLS and two-stage as the data fit is a relaxed numerical solution.

4.1.1. One Parameter Fitzhugh–Nagumo Example

This example highlights the performance when the prior is inconsistent with the likelihood by its over-emphasis of an unimportant local likelihood mode. For ease of visualization, we consider a one parameter model while holding the rest of the parameters in Equations (5) and (6) fixed to the values, $a = 0.2, b = 0.2, \sigma_V^2 = 0.05^2, \sigma_R^2 = 0.05^2, V(0) = -1, R(0) = 1$, with $\boldsymbol{\theta} = c$ being the only parameter to estimate. The 401 evenly spaced observations for each of $V(t)$ and $R(t)$ were simulated with $c = 3$ and are shown in Figure 1. The prior $P(c) = N(14, 2)$ is set up based on the belief that oscillations in the data occur at half the True frequency of oscillation and emphasize a local likelihood mode. Figure 1(a) and (b) show the likelihood and its multiple modes separated by deep valleys measuring several thousands of units on the log scale.

IMIS-ShOpt was run using $Q = 3$ strategies; NLS, two-stage, and GP, from $D = 4$ starting points. To maintain algorithmic comparability, IMIS-Opt was run with $D = 12$ initializations.

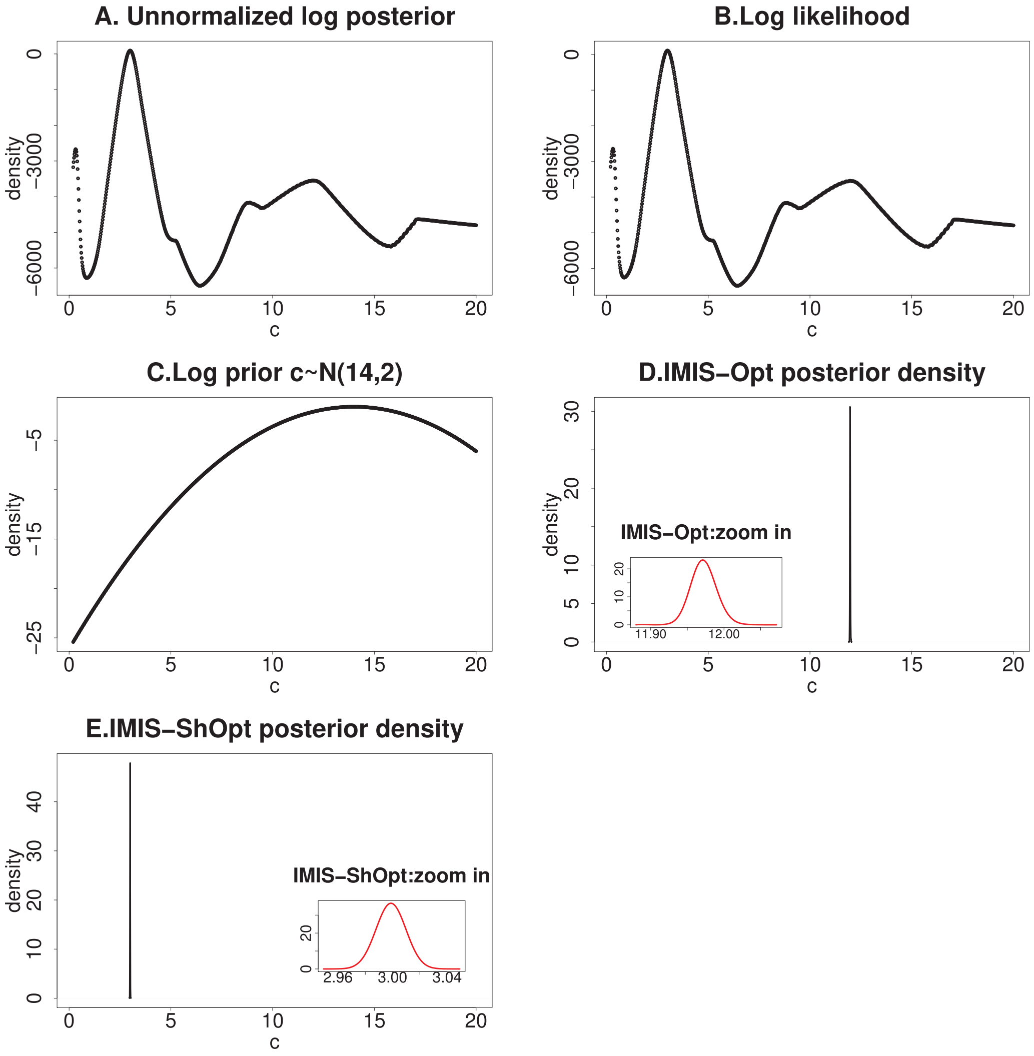

Figure 1. Impact of the disagreement between the likelihood and prior on IMIS-Opt and IMIS-ShOpt posterior estimates in one parameter FhN-ODE model. Log of target distribution (plot A) is obtained by combining the log-likelihood (plot B) with log prior that emphasizes the unimportant local mode (plot C). Densities of the posterior samples obtained from IMIS-Opt and IMIS-ShOpt are presented in plots D and E, respectively. The plots demonstrate that the IMIS-Opt is trapped in the local unimportant mode, while the IMIS-ShOpt is robust to the disagreement between prior and likelihood which leads to discovering the global mode.

Both IMIS-Opt and IMIS-ShOpt were run with $N_0 = 1000$, $B = 100$, and $J = 10,000$, matching the rule of thumb values in Section 3. Random walk and gradient optimization methods are faced with challenges because the basin of attraction of the global likelihood mode begins 4 SD away from the prior mean (Figure 1(a) and (c)). Consequently, IMIS-Opt becomes trapped in the local mode emphasized by the prior (Figure 1(d)). IMIS-ShOpt on the other hand benefits from diversified optimization strategies and explores more widely to find the global mode (Figure 1(e)). Table 1 summarizes the convergence diagnostics for the algorithms, showing that both IMIS-Opt and IMIS-ShOpt have completed.

The $Q = 3$ methods (NLS, two-stage, and GP) combined consistently discover global and local optima. The results from the NLS were highly affected by the initial points, and consequently, were unable to leave the basin of attraction of the local mode. The success of GP in finding the global mode varied (mainly due to choice of λ which was not finely tuned), while the two-stage method proved to be the least sensitive to the initial points and consistently converged to values near the global mode. The exploration of global and local optima is the goal of IMIS-ShOpt, making optimization strategies with mixed results beneficial. In some cases multiple modes are important and incorporating Q optimization strategies allows wider posterior exploration. Figure 2 shows the (un-weighted) importance distribution samples, Θ_D, at the end of the optimization stage of the algorithm. For comparison, Figure 2 includes the target posterior density, which is exceedingly narrow at this horizontal scale. Following the optimization with the importance stages fills in additional samples to refine the importance sampling distribution in important posterior regions.

The natural log Kullback–Leibler divergence between the IMIS-ShOpt sampled distribution and the target posterior are given in Table 2. Sample and target densities are very close together regardless of which is used as the reference. For comparison, IMIS-Opt sampled distribution exhibit much higher natural log Kullback–Leibler divergence from the target posterior compared to that of the IMIS-ShOpt posterior samples. This result is expected because IMIS-Opt misses the global mode. Table 3 shows the CPU time of the algorithm when leaving one optimization strategy out (only using $Q = 2$) to showcase the contribution of each approach to algorithm time. In our case, we

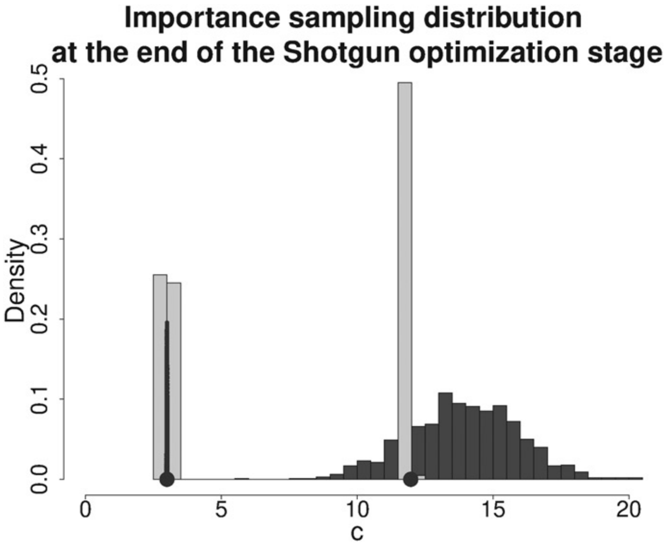

Figure 2. The first two stages (initial and optimization) of building the IMIS-ShOpt importance sampling distribution in one parameter FhN model. Dispersed histogram (dark gray distribution) corresponds to the initial importance sampling distribution. The pointed distributions (light gray color) denote samples obtained from the optimization stage. The dots (dark gray) at $c = 3$ and $c = 12$ represent the global and the unimportant local mode, respectively, discovered by different optimizers. The vertical line at $c = 3$ (dark gray) represents the low variance posterior density from IMIS-ShOpt obtained from the final resampling stage. This plot demonstrates that after finishing the optimization stage, IMIS-ShOpt has already found the most important modes including the global mode, and hence, the importance sampling stage of the IMIS-ShOpt is used for adjusting the weights of the modes.

Table 1. Convergence monitoring for the one parameter FhN and full FhN.

	IMIS-Opt one parameter FhN	IMIS-ShOpt one parameter FhN	IMIS-ShOpt full FhN
The raw marginal likelihood	−5.54	−9.39	−190.26
Expected number of unique points	6323.36	6332.88	6387.48
Maximum weight	0.0002	0.0002	0.0003
Effective sample size	9859.66	9817.47	9449.78
Entropy	0.98	0.98	0.88
Variance	0.18	0.22	3.52

NOTE: Convergence diagnostic monitoring tools from the Section 3.

Table 2. Kullback–Leibler (KL) divergence between target distribution for the one parameter FhN model and the posterior densities obtained from IMIS-ShOpt and IMIS-Opt.

	Target to posterior	Posterior to target
KL divergence, IMIS-ShOpt	0.0016	0.0010
KL divergence, IMIS-Opt	3.381	8.474

want to explore the parameter space widely, so we did not spend much time tuning h in Equation (9) or λ in Equation (11). In both GP and two-stage, the computation speed can be adjusted somewhat with these tuning parameters.

IMIS-ShOpt converges in $k = 103$ iterations. By contrast, the optimization stage in the IMIS-Opt takes $k = 96$ iterations for the algorithm to converge. However, in this case, IMIS-Opt does not converge to the appropriate posterior as it remains trapped in the local mode whereas the ShOpt enables wider exploration and finds the global mode.

The alternative strategy of using an uninformative prior in the FhN example requires a massive increase in the number of initial particles to capture the global mode. The likelihood as shown outlines only a few different modes, however, additional, but much weaker, local modes exist when expanding out the dominant range of the prior to larger values. In this example, IMIS-ShOpt and the ShOpt strategy builds robustness to the practical challenges when selecting a prior which is strongly informed but biased.

Table 3. Time until convergence for IMIS-ShOpt using these $Q = 2$ optimization strategies for the one parameter FhN Model.

	GP and two-stage	NLS and two-stage	NLS and GP
Wall-clock (sec)	323.52	444.51	189.28

4.1.2. Full Fitzhugh–Nagumo Example

In this example, the full FhN model is used with vector of parameters $\boldsymbol{\theta} = \left[a, b, c, \sigma_V^2, \sigma_R^2, V(0), R(0)\right]$. Table 4 presents prior specifications and compares them with the one parameter model of the previous example. The full FhN model has a complex likelihood surface including ripples and ridges along with the multi-modality from the previous section (Ramsay et al. 2007). This example shows the IMIS-Shopt performance with higher complexity compared to the toy one parameter model of Section 4.1.1.

The target posterior is based on the likelihood in Equation (6). Once again the IMIS-ShOpt is set up with $Q = 3$ ($D = 4$, $B = 700$, $N_0 = 7000$, and $J = 10{,}000$); NLS in Equation (7), two-stage in Equation (10), and GP in Equations (11) and (12). In ShOpt, the posterior samples $\hat{\boldsymbol{\theta}}$ are obtained by combining the results from different optimization criteria, while the Hessian matrices evaluated at $\hat{\boldsymbol{\theta}}$ are obtained using the target posterior. Table 1 shows the convergence criteria results for the one parameter FhN and the full FhN (previous section).

Figure 3(b) and (c) demonstrates the cause of the posterior modes in terms of data fit. Although the prior for the parameter c does not adequately cover the global mode, the IMIS-ShOpt recovers the two and a half oscillations of the true trajectories in the one parameter FhN and full FhN. By contrast, the resampled trajectories obtained from the IMIS-Opt (Figure 3(a)), recover only one oscillation of the true trajectories, while missing the other one-and-a-half oscillation. If IMIS-Opt used a stochastic optimizer or an evolutionary optimizer instead of a gradient method, it's possible that the global maximum could have been found.

4.2. Susceptible-Infected-Removed (SIR) Epidemiological Example

In this example, we consider a susceptible-infected-removed (SIR) epidemiological model with a mixture of continuous and discrete parameters. The data is from the second black plague outbreak in the village of Eyam, UK, from June 19, 1666 to November 1, 1666 (Massad et al. 2004). Since the village had been quarantined, the population size is fixed to $N = 261$ and is stratified into states of susceptible $S(t)$, infected $I(t)$, and removed $R(t)$ individuals, $N = S(t) + I(t) + R(t)$. $R(t)$ corresponds to the number of deaths up to time t, because there is no recovery from the plague (Campbell and Lele 2014; Golchi

Table 4. The two FhN models—prior specifications.

	a	b	c	σ_V^2	σ_R^2	$V(0)$	$R(0)$
One parameter FhN	0.2	0.2	$N(14, 2)$	0.05	0.05	-1	1
Full FhN	$N(0, .4)$	$N(0, .4)$	$N(14, 2)$	*IGamma* (3, 3)	*IGamma* (3, 3)	$N(-1, .5)$	$N(1, .5)$
True values	0.2	0.2	3	0.05	0.05	-1	1

NOTE: The two FhN models—in the one parameter FhN, prior has been assigned only for the parameter c, while the rest of the parameters are fixed to their true values. In the full FhN, the prior distributions have been assigned for all parameters.

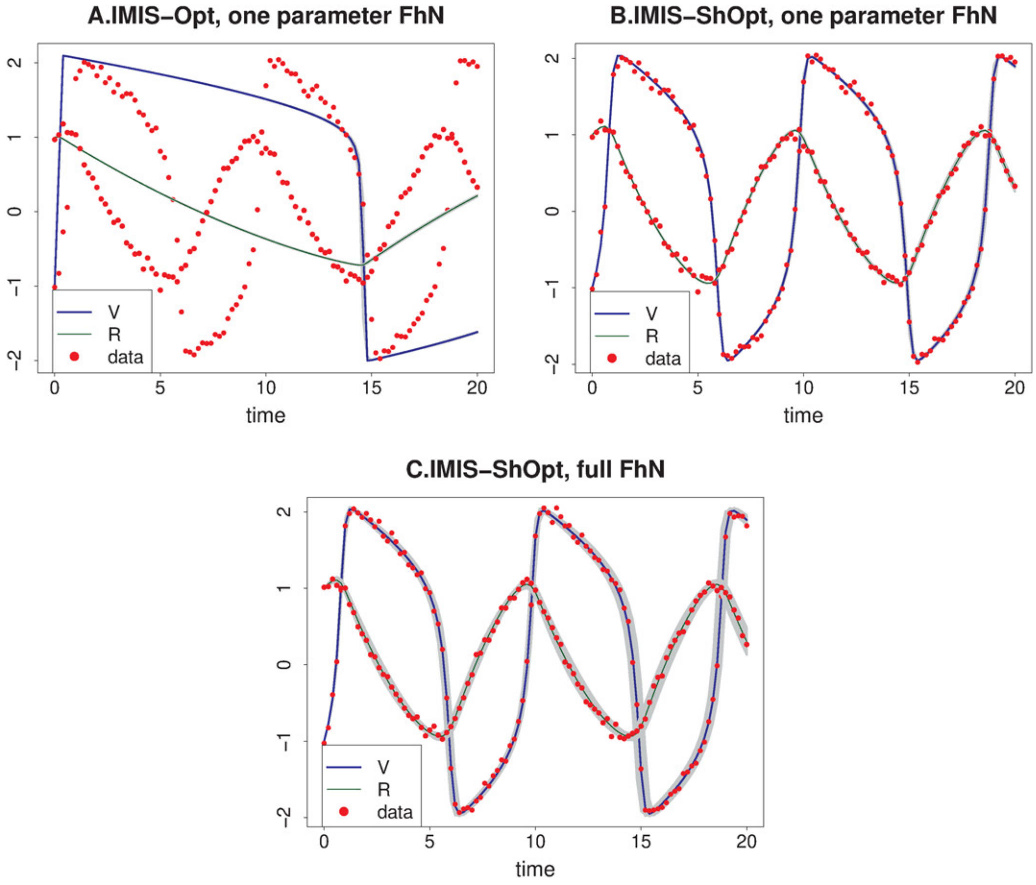

Figure 3. Resampled trajectories obtained from IMIS-Opt and IMIS-ShOpt posterior samples from the FhN model. Plots A and B are generated using posterior samples from one parameter FhN model obtained from IMIS-Opt and IMIS-ShOpt, respectively. Plot C is obtained using posterior samples from the full FhN model from IMIS-ShOpt. The gray lines represent 10,000 resampled trajectories, the solid thick blue and thin green lines correspond to the resampled trajectories at the posterior mean values for the state variables V and R, respectively. The red points represent the data, which were simulated from the vector of true parameters values $\boldsymbol{\theta} = (a = 0.2, b = 0.2, c = 3, V(0) = -1, R(0) = 1)^{'}$. Plots B and C demonstrate that the IMIS-ShOpt finds the global optimum thus, exhibiting a good fit to the data, while plot A shows that the IMIS-Opt gets trapped in the unimportant local mode thus exhibiting suboptimal fit to the data.

and Campbell 2016). The disease spread dynamics are based on the ODE model:

$$\frac{dS}{dt} = -\beta S(t)I(t), \ \frac{dI}{dt} = \beta S(t)I(t) - \alpha I(t), \ \frac{dR}{dt} = \alpha I(t), \tag{13}$$

where α describes the rate of death once the individual is infected and β describes the plague transmission. In order for the ODE system in Equation (13) to be numerically solved, the discrete initial states $S(0), I(0)$, and $R(0)$ are required. Since the number of removed at the initial time is 0, $R(0) = 0$, it follows that $S(0) = N - I(0)$, the initial states of the system reduce to $I(0)$. Hence, parameters of the model are $\boldsymbol{\theta} = (\alpha, \beta, I(0))\,'$. The data $\boldsymbol{Y} = (y_1, \ldots, y_n)'$ comprise of the cumulative number of deaths, $R(t)$, up to times $(t_1, \ldots, t_n)$, $n = 136$. The likelihood of the data followed a binomial distribution with expected value equal to the solution $R_{(\alpha,\beta,I(0))}(t)$ to the system in Equation (13). States S(t) and I(t) are not observed, however, the number of infected at the end of the plague is 0, and the number of infected at time t_{n-1} must, therefore, equal 1 (Campbell and Lele 2014). These two additional data points on number of infected individuals $\boldsymbol{X} = (x_{n-1} = 1, x_n = 0)'$ at times $(t_{n-1}, t_n)'$ were modeled using binomial likelihood with expected value equal to the solution $I_{(\alpha,\beta,I(0))}(t)$ to the system in Equation (13) at $t \in (t_{n-1}, t_n)^{'}$. The resulting likelihood is

$$P(\boldsymbol{Y} \mid \alpha, \beta, I(0)) = \prod_{i=1}^{n} \text{Binomial}\left(y_i \mid N, \frac{R_{(\alpha,\beta,I(0))}(t_i)}{N}\right) \times \prod_{i=n-1}^{n} \text{Binomial}\left(x_i \mid N, \frac{I_{(\alpha,\beta,I(0))}(t_i)}{N}\right). \tag{14}$$

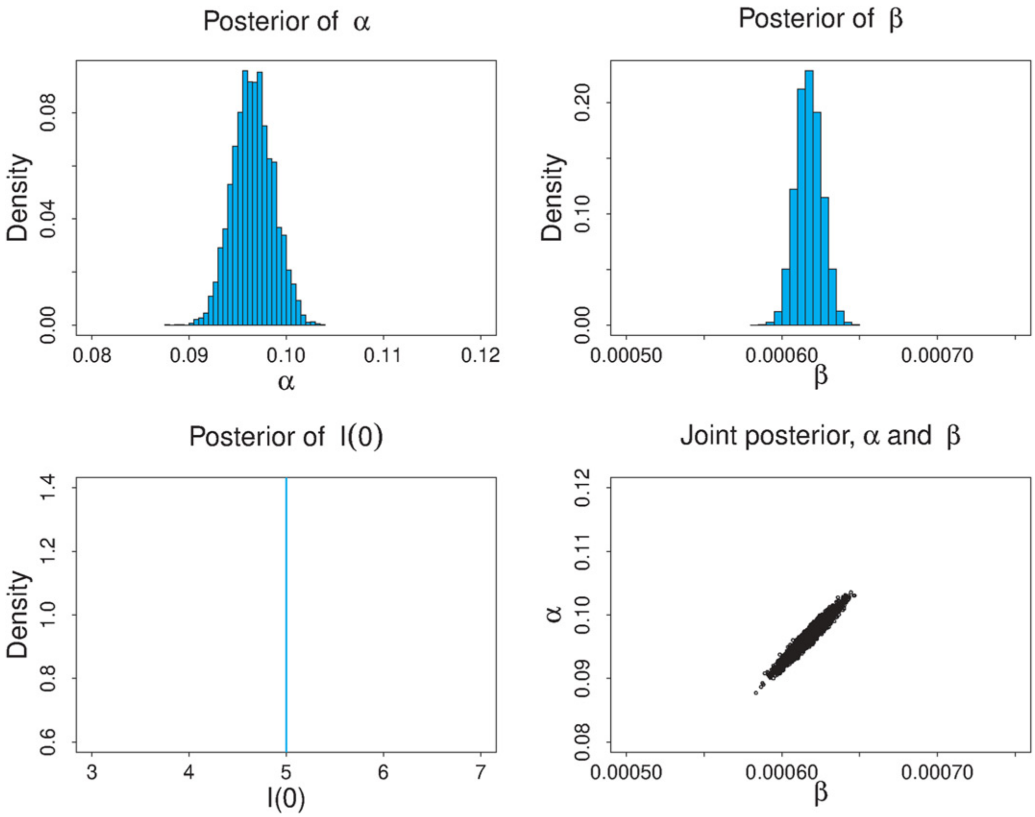

Figure 4. Posterior samples of the SIR-ODE model from IMIS-Opt. Marginal distributions of sampled parameters $\alpha, \beta,$ and $I(0)$, and bivariate joint posterior distribution of α and β are shown. The plots demonstrate that the IMIS-Opt gets trapped in one posterior mode, thus failing to explore the other two important modes.

Prior distributions for $\boldsymbol{\theta} = (\alpha, \beta, I(0))'$ were chosen to be:

$$\alpha, \beta \sim \text{Gamma}(1, 1), I(0) \sim \text{Binomial}\left(N, \frac{5}{N}\right). \quad (15)$$

The challenge of this model is the mixture of discrete and continuous parameters. Consequently, we employ the ShOpt strategy targeting different conditional likelihoods rather than different optimization algorithms. Shotgun optimization applied to the SIR-ODE model uses the $D = 3$ highest weights points to initialize the optimizer, and $Q = 10$ likelihoods conditional on fixed discrete values of $I(0) \in \{1, 2, 3, \ldots, 10\}$. The other IMIS-ShOpt tuning parameters were set to $N_0 = 3000, D = 3, B = 1000, J = 10{,}000$. IMIS-Opt was run with $D = 30$ to maintain a the same number of optimizations and the other parameters set to match IMIS-ShOpt. See Appendix (supplementary materials) for implementation details.

Table 5 shows the computational time in seconds needed to run the IMIS-ShOpt in comparison to that of the IMIS-Opt for the SIR model. The IMIS-Opt is slower than the IMIS-ShOpt, since IMIS-Opt runs the same optimizer 30 times. Figure 5 illustrates multi-modality and topological challenges of the posterior space of the SIR model from the results of IMIS-ShOpt. Marginal distributions of the two continuous parameters α and β exhibit isolated modes. Clouds in the bivariate plot of α and β depict the four modes corresponding to the discrete values of $I(0) = \{6, 5, 4, 3\}$ from left to right. The results of IMIS-ShOpt coincide with those in Campbell and Lele (2014). Although IMIS-Opt finds the global mode, it misses nearby important modes and populates the importance distribution only near the dominant mode (Figure 4). Convergence diagnostics for both algorithms are given in Table 1 in the Appendix (supplementary materials), where both IMIS-Opt and IMIS-ShOpt show convergence to a terminal density. Despite the diagnostic output, without the wider exploration of the parameter space, algorithmic convergence may not coincide with convergence to the target distribution.

Table 5. Computational time in seconds to run IMIS-ShOpt or IMIS-Opt on the SIR model.

IMIS-Opt	IMIS-ShOpt
737.111	416.019

4.3. Parameter Estimation With IMIS-ShOpt Using Synthetic Likelihood

In this section, we introduce the IMIS-ShOpt with synthetic likelihood (Wood 2010) which borrows ideas from the approximate Bayesian computation (ABC) framework. ABC methods (Tavaré et al. 1997; Pritchard et al. 1999) provide a framework for inference in cases where the likelihood is intractable or very costly to evaluate, but simulating data from the model is relatively easy.

Direct likelihood-based inference breaks down in chaotic stochastic systems since small changes in $\boldsymbol{\theta}$ cause drastic changes in the system trajectories and the stochasticity leads to expensive-to-evaluate or intractable, nonsmooth, and

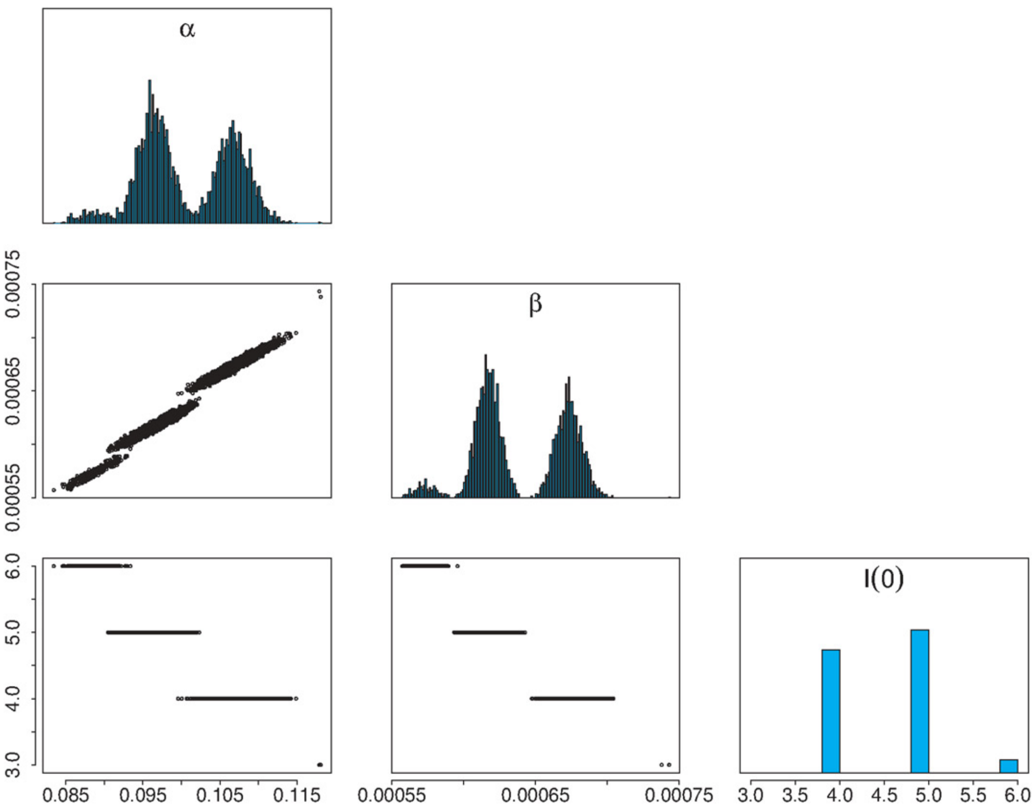

Figure 5. Posterior samples of the SIR-ODE model from IMIS-ShOpt. Marginal (diagonal) and bivariate joint (off-diagonal) posterior distributions of sampled parameters α, β, and $I(0)$ are shown. The plots demonstrate that IMIS-ShOpt successfully explores the three important modes in the posterior space.

chaotic likelihoods (Wood 2010). However, it is relatively easy to simulate data from such models giving access to synthetic likelihood inference. Following Wood (2010), a synthetic likelihood is constructed by comparing summary statistics which capture the important dynamics in the data. Ideally, the summary statistics would be sufficient statistics, however, in practice, they are rarely obtainable. Instead, a set of summary statistics constructed using known dynamics of interest could be used to capture different features of the data. Among the available summaries, those with small variance and high sensitivity to parameter changes are preferred choices.

To avoid the requirement of the tolerance levels and the distance measure needed in ABC, and to gain the efficiency from the ShOpt thereof, we approximate the likelihood function with the synthetic likelihood (Wood 2010). Although the synthetic likelihood approach employs ideas from the ABC framework, the synthetic likelihood behaves like a conventional likelihood in the limit, when the number of simulated datasets approaches infinity, but acts with reduced efficiency because of the lack of sufficient statistics.

Following Wood (2010), the synthetic likelihood can be constructed as follows. For parameters $\boldsymbol{\theta}$, N_Z simulated datasets $\boldsymbol{Z} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{N_Z}\}$ are generated from $P(\boldsymbol{Z} \mid \boldsymbol{\theta})$, and the vector of summary statistics $\boldsymbol{S}(\boldsymbol{Z}) = \{\boldsymbol{s}(\boldsymbol{Z}_1), \ldots, \boldsymbol{s}(\boldsymbol{Z}_{N_Z})\}$ is calculated for each simulated dataset, exactly as the summary statistics $\boldsymbol{S}(\boldsymbol{Y})$ are calculated from the observed data. The mean of the N_Z summary statistics, $\hat{\boldsymbol{\mu}}_\theta = \frac{\sum_{i=1}^{N_Z} \boldsymbol{s}(\boldsymbol{Z}_i)}{N_Z}$, and the variance-covariance matrix, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$, are used to construct the synthetic likelihood as $\text{MVN}(\boldsymbol{S} \mid \hat{\boldsymbol{\mu}}_\theta, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}})$, that is,

$$\mathcal{L}_s(\boldsymbol{\theta} \mid \boldsymbol{S}(\boldsymbol{Y})) = -\frac{1}{2}(\boldsymbol{S}(\boldsymbol{Y}) - \hat{\boldsymbol{\mu}}_\theta)' \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{S}(\boldsymbol{Y}) - \hat{\boldsymbol{\mu}}_\theta) - \frac{1}{2} \log |\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}|. \tag{16}$$

The target likelihood for importance weights is defined over the entire set of available summary statistics.

The theta-Ricker model is a chaotic stochastic discrete time model, where full likelihood-based inference fails, but it is relatively easy to simulate data from the model. Following Gilpin and Ayala (1973), the ecological theta-Ricker model, states that the abundance of the population in the next time point, N_{t+1}, is equal to the abundance at the current time point N_t, multiplied by the exponent of the growth rate, $\exp\left(r(1 - \frac{N_t}{K})^{\tilde{\theta}} + \epsilon_t\right)$, over the time step t. The process noise, also known as environmental noise is modeled as $\epsilon_t \sim N(0, \sigma_p^2)$ and K quantifies carrying capacity. The theta-Ricker model can be written as follows,

$$N_{t+1} = N_t \exp\left(r\left(1 - \left(\frac{N_t}{K}\right)^{\tilde{\theta}}\right) + \epsilon_t\right), \tag{17}$$

The theta-Ricker model is defined with parameters $\boldsymbol{\theta} = [r, \phi, \sigma_p^2, \tilde{\theta}]$. The data are outcomes of the Poisson distribution with mean ϕN_t, where ϕ is a scaling parameter,

$$y_t \sim \text{Poisson}(\phi N_t).$$

```
Algorithm 3 The IMIS-ShOpt with synthetic likelihood
Goal: Parameter estimation
Input: Data, likelihood function, synthetic likelihood function, prior distribution and the model.
Initialize $B$ – the number of incremental points, $D$ – the number of different initial points for the optimization, $Q$ – the number of different optimization criteria, $N_0$ – the number of initial samples from the prior and $J$ – the number of resampled points
Initial stage: Draw $N_0$ samples $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{N_0}\}$ from the prior distribution $P(\boldsymbol{\theta})$, set $k = 0$.
    For each $\boldsymbol{\theta}_i$, $i = 1, \ldots, N_0$, simulate $N_{\boldsymbol{Z}}$ vectors of replicate data $\boldsymbol{Z}_i = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{N_Z}\}$ from the model, $P(\boldsymbol{Z} \mid \boldsymbol{\theta}_i)$.
    For each $\boldsymbol{\theta}_i$, $i = 1, \ldots, N_0$, calculate the vector of entire set of available summary statistics, $\boldsymbol{S}(\boldsymbol{Z}) = \{\boldsymbol{s}(\boldsymbol{Z}_1), \ldots, \boldsymbol{s}(\boldsymbol{Z}_{N_Z})\}$ and construct the synthetic likelihood using Equation (16).
    For each $\boldsymbol{\theta}_i$, $i = 1, \ldots, N_0$ calculate the sampling weights,
        $$w_i^* = \frac{\mathcal{L}_s(\boldsymbol{\theta}_i \mid \boldsymbol{S}(\boldsymbol{Y}))}{\sum_{j=1}^{N_0} \mathcal{L}_s(\boldsymbol{\theta}_j \mid \boldsymbol{S}(\boldsymbol{Y}))} \qquad (18)$$
    Optimization stage:
    for $d = 1 : D$ do
        Find the $d$th maximum weight point $\boldsymbol{\theta}_d^{(\text{initial})} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \boldsymbol{w}^*(\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}_{d-1}$ to initialize Q optimizers.
        for $q = 1 : Q$ do
            Randomly select summary statistics $\tilde{\boldsymbol{S}}_{\boldsymbol{dq}}(\boldsymbol{Y})$ and obtain $\boldsymbol{\theta}_{d,q}^{(\text{Opt})} = \arg\max_{\boldsymbol{\theta}} \tilde{\mathcal{L}}_{dq}\left(\boldsymbol{\theta} \mid \tilde{\boldsymbol{S}}_{\boldsymbol{dq}}(\boldsymbol{Y})\right)$. Obtain the corresponding inverse Hessian $\boldsymbol{\Sigma}_{d,q}^{(\text{Opt})}$ from $\mathcal{L}_s(\boldsymbol{\theta} \mid \boldsymbol{S}(\boldsymbol{Y}))$.
            Update $\boldsymbol{\Theta}_d$ by excluding $\frac{N_0}{DQ}$ nearest neighbor points, $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_{d-1}$, that minimize the Mahalanobis distance,
                $$(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{d,q}^{(\text{Opt})})'(\boldsymbol{\Sigma}_{d,q}^{(\text{Opt})})^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{d,q}^{(\text{Opt})}). \qquad (19)$$

Algorithm 3 *
Algorithm 3 The IMIS-ShOpt with synthetic likelihood - continued
            Draw $B$ samples $\boldsymbol{\theta}_{1:B} \sim \text{MVN}(\boldsymbol{\theta}_{\text{d,q}}^{(\text{Opt})}, \boldsymbol{\Sigma}_{\text{d,q}}^{(\text{Opt})})$; add these points to the importance sampling distribution $P(\boldsymbol{\theta} \mid \boldsymbol{Y})$ and evaluate $H_k = \text{MVN}(\boldsymbol{\theta}_{1:\text{B}} \mid \boldsymbol{\theta}_{\text{d,q}}^{(\text{Opt})}, \boldsymbol{\Sigma}_{\text{d,q}}^{(\text{Opt})})$.
        end for
    end for
    Importance sampling stage:
    For each $\boldsymbol{\theta}_i, i = 1, \ldots, N_k$ calculate weights:
        $$w_i^{(k)} = \frac{cP(\boldsymbol{\theta}_i)\mathcal{L}_s(\boldsymbol{\theta}_i \mid \boldsymbol{S}(\boldsymbol{Y}))}{\frac{N_0}{N_k}P(\boldsymbol{\theta}_i) + \frac{B}{N_k}\sum_{s=1}^{k} H_s(\boldsymbol{\theta}_i)}, \qquad (20)$$
    where $N_k = N_0 + B(QD + k)$ and $c = 1/\sum_{i=1}^{N_k} w_i^{(k)}$.
    while $\sum_{1}^{N_k}(1 - (1 - w^{(k)})^J) < J(1 - \exp(-1))$, that is, importance sampling weights are not approximately uniform
    do
        Update $k = k + 1$.
        Choose a maximum weight input, $\boldsymbol{\theta}_k$, and estimate $\boldsymbol{\Sigma}_k$ as the weighted covariance of B inputs with smallest Mahalanobis distance,
            $$w_p(\boldsymbol{\theta})\,(\boldsymbol{\theta} - \boldsymbol{\theta}_k)'\,(\boldsymbol{\Sigma}_{\boldsymbol{\pi}})^{-1}\,(\boldsymbol{\theta} - \boldsymbol{\theta}_k),$$
        where the weights are $w_p(\boldsymbol{\theta}) = c_1(\boldsymbol{w}^{(k-1)} + 1/N_k)$, $\Sigma_\pi$ is the covariance of the initial importance distribution and $c_1 = 1/w_p(\boldsymbol{\theta})$.
        Draw $B$ samples $\boldsymbol{\theta}_{1:B} \sim \text{MVN}(\boldsymbol{\theta}_\text{k}, \boldsymbol{\Sigma}_\text{k})$; add these points to the importance sampling distribution and evaluate $H_k = \text{MVN}(\boldsymbol{\theta}_{1:\text{B}} \mid \boldsymbol{\theta}_{\text{m,k}}, \boldsymbol{\Sigma}_\text{k})$.
        Calculate weights $w^{(k)}$ using Equation (20).
    end while
    Resampling stage:
    Resample $J$ points with replacement from $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{N_k}\}$ and weights $w^{(k)}$.
```

The IMIS-ShOpt algorithm was used to estimate the parameters of the theta-Ricker model. The data were simulated from $\boldsymbol{\theta} = \left[\log r = 0.5, \phi = 4, \sigma^2 = 0.01, \log\tilde{\theta} = 1\right]$ at T=50 time steps with initial population $N_0 = 3$ and $K = 100$. Prior distributions were defined independently, $\log r \sim N(0.5, 1), \phi \sim \chi^2(df = 4), \sigma_p^2 \sim I\Gamma(\text{shape} = 2, \text{scale} = 0.05), \log\tilde{\theta} = N(1, 1)$.

The set of summary statistics used in IMIS-ShOpt is a modification of the set from Golchi and Campbell (2016),

$$\begin{aligned}\boldsymbol{S}(\boldsymbol{Y}) = \Bigg\{&\text{median}(\boldsymbol{Y}), \sum_{i=1}^{n}\frac{y_i}{n}, \frac{\sum_{i=1}^{n} y\mathbb{I}_{(1,\infty)}(y_i)}{\sum_{i=1}^{n}\mathbb{I}_{(1,\infty)}(y_i)},\\ &\sum_{i=1}^{n} y\mathbb{I}_{(10,\infty)}(y_i), \sum_{i=1}^{n}\mathbb{I}_0(y_i),\\ &\text{Quantile}_{0.75}(\boldsymbol{Y}), \max(\boldsymbol{Y}), \sum_{i=1}^{n}\mathbb{I}_{(100,\infty)}(y_i),\\ &\sum_{i=1}^{n}\mathbb{I}_{(300,\infty)}(y_i),\\ &\sum_{i=1}^{n}\mathbb{I}_{(500,\infty)}(y_i), \sum_{i=1}^{n} y\mathbb{I}_{(800,\infty)}(y_i)\Bigg\}. \qquad (21)\end{aligned}$$

IMIS-ShOpt was performed with $B = 1000, J = 3000, N_0 = 10{,}000, N = 500, D = 4, Q = 3, N_Z = 30$. At each of the $D = 4$ iterations of the optimization stage, $Q = 3$ objective functions targeted random approximations to the synthetic likelihood $\tilde{\mathcal{L}}_{dq}\left(\boldsymbol{\theta} \mid \tilde{\boldsymbol{S}}_{\boldsymbol{dq}}(\boldsymbol{Y})\right)$ where $\tilde{\boldsymbol{S}}_{\boldsymbol{dq}}(\boldsymbol{Y})$ is the $(dq)^{\text{th}}$ random selection of 7 summaries from the 11 possibilities in Equation (21), $\tilde{\boldsymbol{S}}(\boldsymbol{Y}) = \{s_i, s_j, s_k, s_l, s_m, s_o, s_p \mid i, j, k, l, m, o, p \in 1, \ldots, 11\} \subseteq \boldsymbol{S}(\boldsymbol{Y})$. There were $\binom{11}{7} = 330$ possible objective functions that the $DQ = 12$ optimizers could target. These approximations to the target synthetic likelihood explore different regions of the posterior space, and therefore, increase the chances of discover-

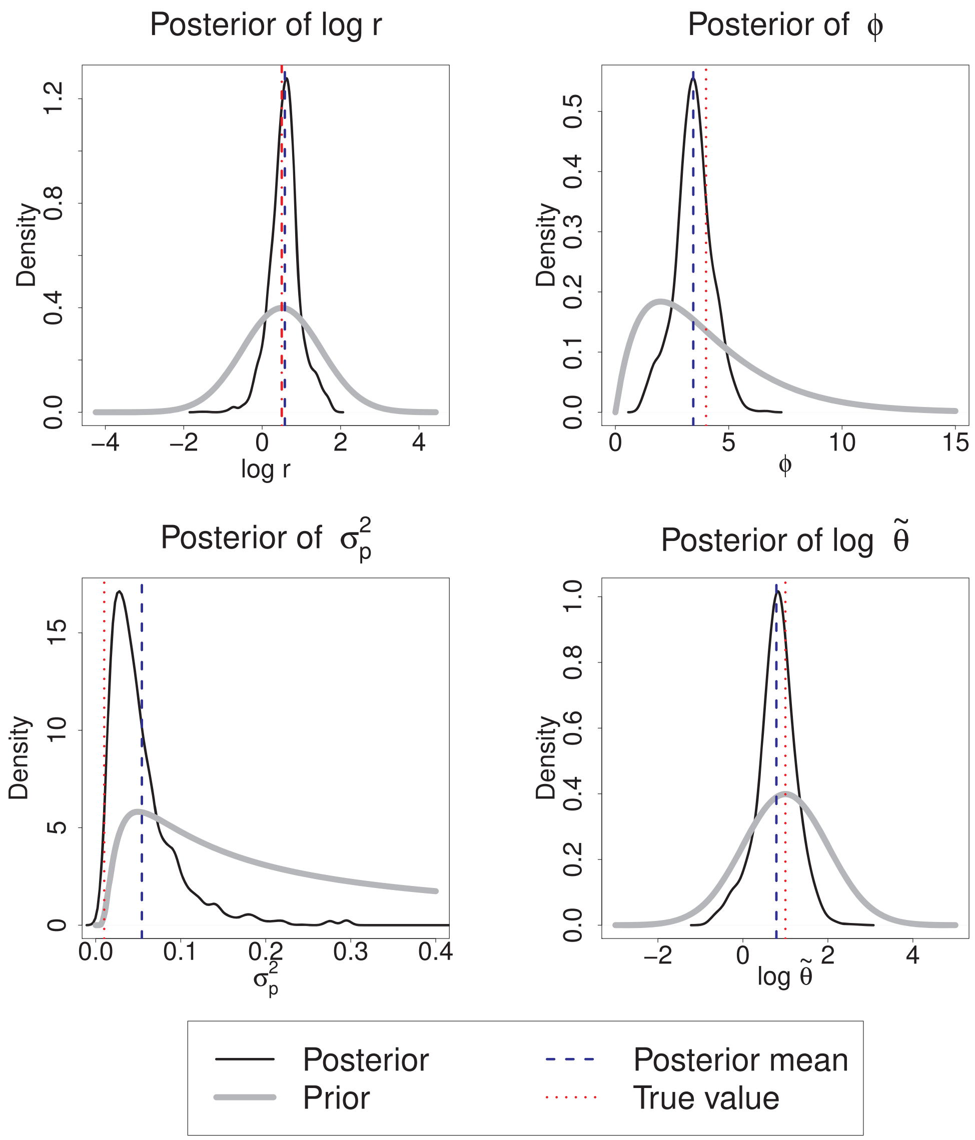

Figure 6. Marginal posterior distributions of the sampled parameters log r, ϕ, σ_p^2, and log $\tilde{\theta}$ obtained from the final resampling stage of the IMIS-ShOpt with synthetic likelihood from the theta-Ricker model. The vertical lines are drawn at the posterior mean (blue dashed) and the true value (red dotted). The thick gray distributions represent the priors. The plots indicate that IMIS-ShOpt successfully explores the posterior space of the theta-Ricker model.

ing additional important posterior modes. The pseudo-code of the IMIS-ShOpt algorithm with synthetic likelihood is given in Algorithm 3.

The idea of optimizing a random subset of the model is common in fitting massive neural network models in Deep Learning. The strategy dubbed "dropout" randomly removes portions of the model (for us a subset of summaries) at each optimization step in order to avoid overfitting and to ease optimization (Goodfellow, Bengio, and Courville 2016). Approximations to the target synthetic likelihood constructed by randomly chosen subsets of summary statistics, are less constrained by information and therefore, more diffuse then the target synthetic likelihood, making them easier to optimize while allowing wider exploration of the target posterior.

Table 6 shows convergence diagnostics. The results, presented as kernel density estimates of the approximate marginal posteriors, are given in Figure 6. Figure 7 shows that the weights of all the particles in the importance sampling distribution before the final resampling stage are nonzero in the neighborhood of the true parameter values. In addition, Figure 7 shows that before the final resampling stage the importance sampling distribution of the process noise variance, σ_p^2, contains particles with negative values, however, these have zero weights and are not resampled in the final stage.

Table 6. Convergence monitoring for the Ricker model.

	Ricker model
The raw marginal likelihood	−10.76
Expected number of unique points	1908.90
Maximum weight	0.003
Effective sample size	2051.11
Entropy	0.78
Variance	21.91

NOTE: Convergence diagnostic monitoring tools from the Section 3.

5. Discussion

This article proposes an importance sampling strategy based on ShOpt, which exploits the NFL theorem for optimization by

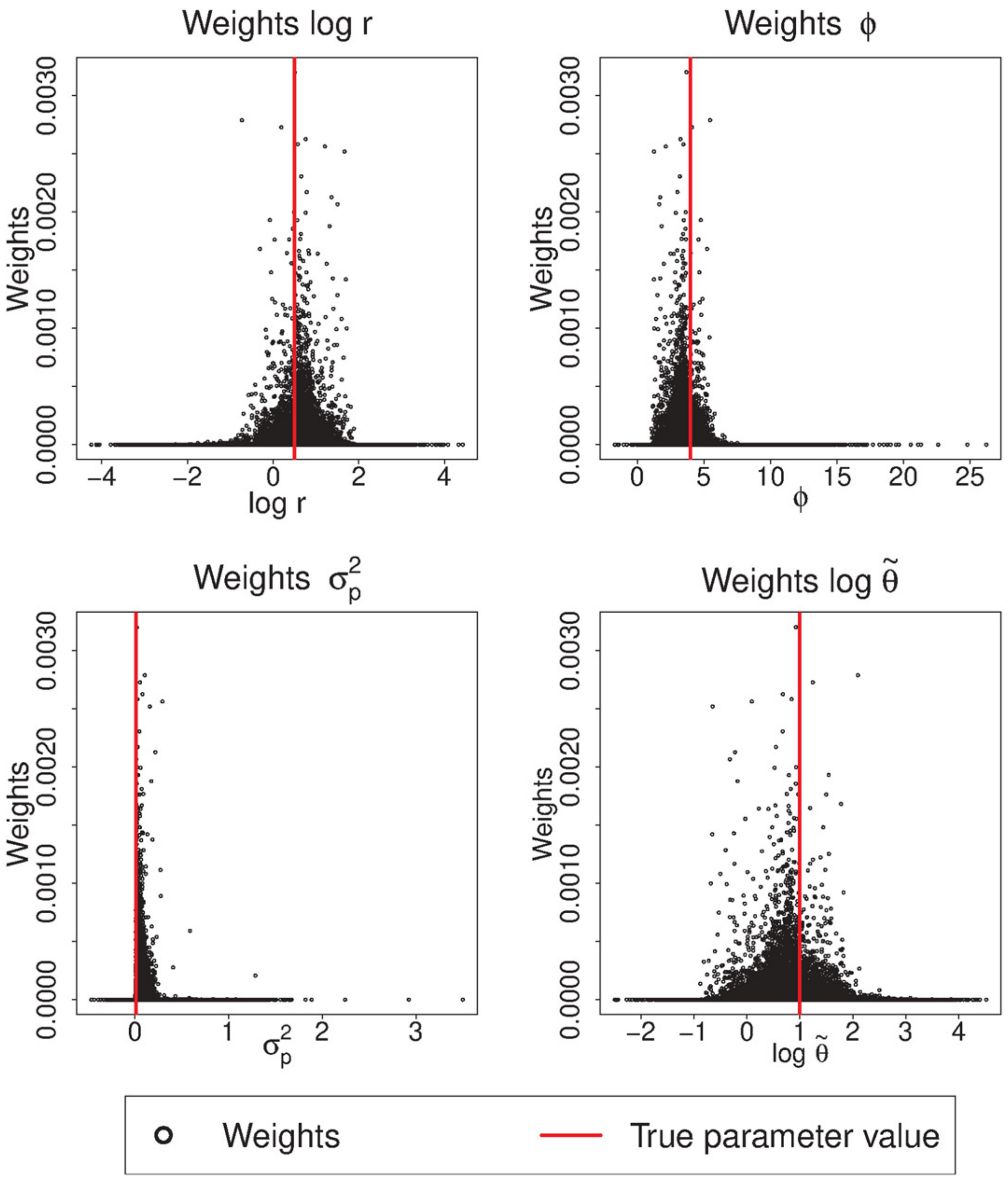

Figure 7. Weights of the particles in the importance sampling distributions of the parameters log r, ϕ, σ_p^2, and log $\tilde{\theta}$ before the final resampling stage for the theta-Ricker model from the IMIS-ShOpt with synthetic likelihood. The vertical lines are drawn at the true parameter values. The plots demonstrate that the weights of the points are concentrated around the true values.

employing different model variations (Sections 4.1.1 and 4.1.2), optimizing with different conditions (Section 4.2), or different optimization criteria (Section 4.3). The solutions we proposed were customized to the modeling scenario. An alternative strategy would be to use a variety of optimizers for a fixed model, for example, stochastic gradient descent, simplex, and higher order methods as the Q optimizers. However, this will not be as effective at increasing the probability of finding a global optimum as exploiting the model structure in the optimizer (Wolpert and Macready 1997).

While IMIS-ShOpt was run with the same number of optimization initializations as IMIS-Opt, the diversity of results attainable from different optimizers provides robustness to cases where the prior and likelihood are inconsistent, the posterior contains distant but important modes, or a single optimizer may computationally struggle in portions of the parameter space. The alternative strategy of selecting a diffuse or uninformative prior for the parameter requires a massive increase in the number of initial samples, especially as the dimension of the problem increases. Instead of introducing philosophical challenges from altering the prior to ease optimization rather than capturing expert opinion, ShOpt provides a means of efficiently targeting resources to important, potentially unexpected posterior regions.

Although IMIS originally developed for estimating marginal likelihoods $P(Y) = \int P(\theta, Y)d\theta$ for model selection, in this article, we focus on parameter estimation to elucidate the benefit of exploiting the NFL theorem. Convergence diagnostics for IMIS and AMIS types of algorithms only describe convergence of the process used to add more samples and more distributions. Without wider exploration of the posterior space convergence of an algorithm may not equate to convergence to the target posterior. The same has been observed for MCMC (Cowles and Carlin 1996). Just as recent MCMC variants designed to exploit wider targeted exploration add robustness to problems caused by challenging posterior topologies, so does Shotgun approach add robustness to complex problems within the importance sampling framework.

Supplementary Materials

Computer Code: Examples in this article are implemented in (R Core Team 2017), using following packages: utils, graphics, parallel, CollocInfer (Hooker, Ramsay, and Xiao 2016), deSolve (Soetaert, Petzoldt, and Setzer 2010), lokern (For R and Enhanced by Martin Maechler 2016), IMIS (Raftery and Le Bao 2012), gtools (Warnes, Bolker, and Lumley 2015), MCMCpack (Martin, Quinn, and Park 2011), mvtnorm (Genz and Bretz 2009), truncnorm (Mersmann et al. 2018), optimx (Nash 2014), coda (Plummer et al. 2006), matrixcalc (Novomestky 2012), numDeriv (Gilbert and Varadhan 2016), flexmix (Grün and Leisch 2008). R code of the implemented examples can be found on github *https://github.com/BiljanaJSJ/IMIS-ShOpt*.

Acknowledgments

The authors thank Dr. Luke Bornn, Dr. Derek Bingham, Dr. Liangliang Wang, and Dr. Russell Steele for the constructive discussions, and Dr. Michael Jack Davis for proof reading the article. The authors are very grateful to the two anonymous referees for their constructive comments.

Funding

This work was supported by grant from Natural Sciences and Engineering Research Council of Canada (RGPIN04040-2014) and the Department of Statistics and Actuarial Sciences at Simon Fraser University.

References

Alkema, L., Raftery, A. E., and Clark, S. J. (2007), "Probabilistic Projections of HIV Prevalence Using Bayesian Melding," *The Annals of Applied Statistics*, 1, 229–248. [807]

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., Pelletier, F., Buettner, T., and Heilig, G. K. (2011), "Probabilistic Projections of the Total Fertility Rate for All Countries," *Demography*, 48, 815–839. [806]

Bacharoglou, A. G. (2010), "Approximation of Probability Distributions by Convex Mixtures of Gaussian Measures," *Proceedings of the American Mathematical Society*, 138, 2619–2628. [806]

Bates, D. M., and Watts, D. G. (1988), *Nonlinear Regression Analysis and Its Applications*, Hoboken, NJ: Wiley. [809,810]

Berger, J. O., Liseo, B., and Wolpert, R. L. (1999), "Integrated Likelihood Methods for Eliminating Nuisance Parameters," *Statistical Science*, 14, 1–28. [807]

Brunel, N. J. (2008), "Parameter Estimation of ODE's Via Nonparametric Estimators," *Electronic Journal of Statistics*, 2, 1242–1267. [809,810]

Campbell, D., and Lele, S. (2014), "An ANOVA Test for Parameter Estimability Using Data Cloning With Application to Statistical Inference for Dynamic Systems," *Computational Statistics & Data Analysis*, 70, 257–267. [812,813,814]

Campbell, D., and Steele, R. (2012), "Smooth Functional Tempering for Nonlinear Differential Equation Models," *Statistics and Computing*, 22, 429–443. [806,810]

Cornuet, J., Marin, J.-M., Mira, A., and Robert, C. P. (2012), "Adaptive Multiple Importance Sampling," *Scandinavian Journal of Statistics*, 39, 798–812. [809]

Cowles, M. K., and Carlin, B. P. (1996), "Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review," *Journal of the American Statistical Association*, 91, 883–904. [818]

Del Moral, P., Doucet, A., and Jasra, A. (2006), Sequential Monte Carlo Samplers," *Journal of the Royal Statistical Society*, Series B, 68, 411–436. [806]

Ding, A. A., and Wu, H. (2014), "Estimation of Ordinary Differential Equation Parameters Using Constrained Local Polynomial Regression," *Statistica Sinica*, 24, 1613. [810]

Douc, R., Guillin, A., Marin, J.-M., and Robert, C. P. (2007), "Convergence of Adaptive Mixtures of Importance Sampling Schemes," *The Annals of Statistics*, 35, 420–448. [809]

Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66* (Vol. 66), Boca Raton, FL: CRC Press. [810]

FitzHugh, R. (1961), "Impulses and Physiological States in Theoretical Models of Nerve Membrane," *Biophysical Journal*, 1, 445. [809]

For R and Enhanced by Martin Maechler (2016), *lokern: Kernel Regression Smoothing with Local or Global Plug-in Bandwidth*, R package version 1.1-8. [819]

Friedman, J., Hastie, T., and Tibshirani, R. (2001), *The Elements of Statistical Learning* (Vol. 1). Berlin: Springer Series in Statistics Springer. [807]

Genz, A., and Bretz, F. (2009), *Computation of Multivariate Normal and t-Probabilities*, Lecture Notes in Statistics. Heidelberg: Springer-Verlag. [819]

Geyer, C. (1991), "Markov Chain Monte Carlo Maximum Likelihood," in *Computing Science and Statistics*, Fairfax Station, VA: Interface Foundation, pp. 156–163. [806]

Geyer, C. J., and Thompson, E. A. (1995), "Annealing Markov Chain Monte Carlo With Applications to Ancestral Inference," *Journal of the American Statistical Association*, 90, 909–920. [806]

Gilbert, P., and Varadhan, R. (2016), *numDeriv: Accurate Numerical Derivatives*, R package version 2016.8-1. [819]

Gilpin, M. E., and Ayala, F. J. (1973), "Global Models of Growth and Competition," *Proceedings of the National Academy of Sciences*, 70, 3590–3593. [815]

Golchi, S., and Campbell, D. A. (2016), "Sequentially Constrained Monte Carlo," *Computational Statistics & Data Analysis*, 97, 98–113. [813,816]

Goodfellow, I., Bengio, Y., and Courville, A. (2016), *Deep Learning*, Cambridge, MA: MIT Press. [817]

Grün, B., and Leisch, F. (2008), "FlexMix Version 2: Finite Mixtures With Concomitant Variables and Varying and Constant Parameters," *Journal of Statistical Software*, 28, 1–35. [819]

Hesterberg, T. (1995), "Weighted Average Importance Sampling and Defensive Mixture Distributions," *Technometrics*, 37, 185–194. [807,809]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–401. [807]

Hooker, G., Ramsay, J. O., and Xiao, L. (2016), "CollocInfer: Collocation Inference in Differential Equation Models," *Journal of Statistical Software*, 75, 1–52. [819]

Hukushima, K., and Nemoto, K. (1996), "Exchange Monte Carlo Method and Application to Spin Glass Simulations," *Journal of the Physical Society of Japan*, 65, 1604–1608. [806]

Kuhn, H., and Tucker, A. (1951), *Proceedings of 2nd Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press. [807]

Liang, H., and Wu, H. (2008), "Parameter Estimation for Differential Equation Models Using a Framework of Measurement Error in Regression Models," *Journal of the American Statistical Association*, 103, 1570–1583. [809,810]

Madigan, D., and Raftery, A. E. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 1535–1546. [807]

Marin, J.-M., Pudlo, P., and Sedki, M. (2012), "Consistency of the Adaptive Multiple Importance Sampling," arXiv preprint arXiv:1211.2548. [809]

Marinari, E., and Parisi, G. (1992), "Simulated Tempering: A New Monte Carlo Scheme," *EPL (Europhysics Letters)*, 19, 451. [806]

Martin, A. D., Quinn, K. M., and Park, J. H. (2011), "MCMCpack: Markov Chain Monte Carlo in R," *Journal of Statistical Software*, 42, 22. [819]

Massad, E., Coutinho, F., Burattini, M., and Lopez, L. (2004), "The Eyam Plague Revisited: Did the Village Isolation Change Transmission From Fleas to Pulmonary?," *Medical Hypotheses*, 63, 911–915. [812]

Mendes-Moreira, J., Soares, C., Jorge, A. M., and Sousa, J. F. D. (2012), "Ensemble Approaches for Regression: A Survey," *ACM Computing Surveys (CSUR)*, 45, 10. [807]

Mersmann, O., Trautmann, H., Steuer, D., and Bornkamp, B. (2018), *truncnorm: Truncated Normal Distribution*, R package version 1.0-8. [819]

Miettinen, K. (2012), *Nonlinear Multiobjective Optimization* (Vol. 12), Berlin: Springer Science and Business Media. [807]

Montgomery, J. M., Hollenbach, F. M., and Ward, M. D. (2012), "Improving Predictions Using Ensemble Bayesian Model Averaging," *Political Analysis*, 20, 271–291. [807]

Nagumo, J., Arimoto, S., and Yoshizawa, S. (1962), "An Active Pulse Transmission Line Simulating Nerve Axon," *Proceedings of the IRE*, 50, 2061–2070. [809]

Nash, J. C. (2014), "On Best Practice Optimization Methods in R," *Journal of Statistical Software*, 60, 1–14. [819]

Novomestky, F. (2012), *matrixcalc: Collection of Functions for Matrix Calculations*, R package version 1.0-3. [819]

Owen, A., and Zhou, Y. (2000), "Safe and Effective Importance Sampling," *Journal of the American Statistical Association*, 95, 135–143. [809]

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006), "Coda: Convergence Diagnosis and Output Analysis for MCMC," *R News*, 6, 7–11. [819]

Poole, D., and Raftery, A. E. (2000), Inference for deterministic simulation models: the "Bayesian Melding Approach," *Journal of the American Statistical Association*, 95, 1244–1255. [806,807]

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999), "Population Growth of Human y Chromosomes: A Study of y Chromosome Microsatellites," *Molecular Biology and Evolution*, 16, 1791–1798. [814]

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [819]

Raftery, A. E., and Bao, L. (2012), *IMIS: Incremental Mixture Importance Sampling*, R package version 0.1. [819]

——— (2010), "Estimating and Projecting Trends in HIV/AIDS Generalized Epidemics Using Incremental Mixture Importance Sampling," *Biometrics*, 66, 1162–1173. [806,807,808,809]

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007), "Parameter Estimation for Differential Equations: A Generalized Smoothing Approach," *Journal of the Royal Statistical Society*, Series B, 69, 741–796. [809,812]

Rubin, D. B. (1987), "The Calculation of Posterior Distributions by Data Augmentation: Comment: A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm," *Journal of the American Statistical Association*, 82, 543–546. [806,807]

——— (1988), "Using the SIR Algorithm to Simulate Posterior Distributions," *Bayesian Statistics*, 3, 395–402. [806,807]

Sbert, M., Havran, V., and Szirmay-Kalos, L. (2018), "Multiple Importance Sampling Revisited: Breaking the Bounds," *EURASIP Journal on Advances in Signal Processing*, 2018, 15. [809]

Seber, G. A. F., and Wild, C. J. (1989), *Nonlinear Regression*, Hoboken, NJ: Wiley. [809]

Soetaert, K., Petzoldt, T., and Setzer, R. W. (2010), "Solving Differential Equations in R: Package Desolve," *Journal of Statistical Software*, 33, 1–25. [819]

Steele, R. J., Raftery, A. E., and Emond, M. J. (2006), "Computing Normalizing Constants for Finite Mixture Models via Incremental Mixture Importance Sampling (IMIS)," *Journal of Computational and Graphical Statistics*, 15, 712–734. [807]

Swendsen, R. H., and Wang, J.-S. (1986), "Replica Monte Carlo Simulation of Spin-Glasses," *Physical Review Letters*, 57, 2607. [806]

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997), "Inferring Coalescence Times From DNA Sequence Data," *Genetics*, 145, 505–518. [814]

Varah, J. (1982), "A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations," *SIAM Journal on Scientific and Statistical Computing*, 3, 28–46. [809,810]

Walley, P., and Moral, S. (1999), "Upper Probabilities Based Only on the Likelihood Function," *Journal of the Royal Statistical Society*, Series B, 61831–847. [807]

Warnes, G. R., Bolker, B., and Lumley, T. (2015), *gtools: Various R Programming Tools*, R package version 3.5.0. [819]

Wolpert, D. H., and Macready, W. G. (1997), "No Free Lunch Theorems for Optimization," *IEEE Transactions on Evolutionary Computation*, 1, 67–82. [806,818]

Wood, S. N. (2010), "Statistical Inference for Noisy Nonlinear Ecological Dynamic Systems," *Nature*, 466, 1102–1104. [814,815]

Zhang, C., and Ma, J. (2008), "Comparison of Sampling Efficiency Between Simulated Tempering and Replica Exchange," *The Journal of Chemical Physics*, 129, 134112. [806]