Efficient parameter estimation for complex simulation-based generative models

Associate Professor Christopher Drovandi

School of Mathematical Sciences
ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS)
Queensland University of Technology

15th October

Outline

Bayesian Statistics

In Bayesian statistics we are interested in sampling from the posterior:

$$p(\theta|y) \propto p(y|\theta)p(\theta),$$

where $p(y|\theta)$ is the likelihood and $p(\theta)$ is the prior.

Markov Chain Monte Carlo

Construct ergodic Markov chain with invariant distribution $p(\theta|y)$ (Metropolis et al., 1953)

A common MCMC algorithm is Metropolis Hastings (MH) MCMC, where proposals θ^* are accepted with probability

$$\min\left(1, \frac{p(y|\theta^*)p(\theta^*)q(\theta|\theta^*)}{p(y|\theta)p(\theta)q(\theta^*|\theta)}\right),$$

where $q(\cdot)$ is the proposal density.

For complex models, $p(y|\theta)$ may not be computationally feasible.

Approximate Bayesian Computation

Approximate Bayesian computation (ABC[1]) is current state-of-the-art likelihood-free Bayesian method.

ABC prefers parameter values that simulates x close to y, in terms of summary statistics $S(\cdot)$.

Targets posterior conditional on observed summary $p(\theta|s_y) \propto p(s_y|\theta)p(\theta)$ where $s_y = S(y)$.

Estimates $p(s_y|\theta)$ non-parametrically via simulation[2].

Choice of summary function $S(\cdot)$ trade-off between information loss and dimensionality.

[1] Handbook of Approximate Bayesian Computation. Eds: Sisson, Fan, Beaumont. CRC Press.

[2] Blum 2010. Approximate Bayesian Computation: a nonparametric perspective. JASA.

Approximate Bayesian Computation

ABC Approximation of likelihood $p(s_y|\theta)$

- Simulate n iid datasets, denoted $x_{1:n} = (x_1, ..., x_n)$, from the model based on θ.
- Calculate n sets of summary statistics, $s_{1:n} = (s_1, ..., s_n)$
- The intractable $p(s_y|\theta)$ is replaced with the estimated ABC likelihood,

$$\hat{p}_\epsilon(s_y|\theta) = \frac{1}{n}\sum_{i=1}^{n} K_\epsilon\left(\rho(s_y, s_i)\right).$$

  - $\rho(\cdot)$ is called the discrepancy function
  - $K_\epsilon(\cdot)$ is a kernel weighting function with bandwidth ϵ
  - ϵ is called the ABC tolerance (bias/variance trade-off)

Approximate Bayesian Computation

Disadvantages

- Highly sensitive to choice of tuning parameter ϵ, $\rho(\cdot)$ and to a lesser extent $K_\epsilon(\cdot)$
- No standard way to select ϵ or $\rho(\cdot)$.
- Suffers from curse of dimensionality with respect to size of summary statistic

Motivating Example - Fowler's Toads

Individual-based model[1] of a species called Fowler's Toads (*Anaxyrus fowleri*).

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 ($n_t = 66$ and $n_d = 63$ here). Denote matrix $\boldsymbol{Y}$.

$\boldsymbol{Y}$ is summarised to 4 sets comprising the moving distances for time lags of $1, 2, 4, 8$ days.

[1] Marchand et al 2017. A stochastic movement model reproduces patterns of site fidelity and long-distance dispersal in a population of Fowler's toads. Ecological Modelling.

Fowler's Toads – Model

For each toad, first generate an overnight displacement, Δy. Displacement assumed to follow Lévy-alpha stable distribution family with parameters α and η.

Total returning probability is a constant p_0.

- Model 1: Random return to site.
- Model 2: Nearest return
- Model 3: Probability of return to site depends on distance (extra parameter d_0). Not considered in this talk.

Parameter of interest $\theta = (\alpha, \eta, p_0)$. Priors from Marchand et al: $\alpha \sim U(1,2)$, $\eta \sim U(10,100)$ and $p_0 \sim U(0,1)$.

Fowler's Toads – Summary Statistic Selection

Number of returns for all four time lags (defined as distance < 10m).

For the non-returns we consider log difference between adjacent p-quantiles with $p = 0, 0.1, \ldots, 1$ and also the median. Repeat for each time lag.

Roughly 40 statistics. Difficult for conventional ABC to deal with.

Parametric Alternatives

It might be possible to overcome some drawbacks of ABC by using a parametric approximation to $p(s_y|\theta)$ instead of the non-parametric approximation used by ABC.

Bayesian Synthetic Likelihood

The synthetic likelihood (SL) method[1] uses a multivariate normal approximation: $p(s_y|\theta) \approx \mathcal{N}(s_y; \mu(\theta), \Sigma(\theta))$.

- Suitable when summary statistics are subject to the central limit theorem
- Transformations to multivariate normality of summary statistics
- Summary statistics from indirect inference
- Popular & convenient choice

We developed a Bayesian version of synthetic likelihood called BSL[2].

[1] Wood 2010. Statistical inference for noisy nonlinear ecological dynamic systems. Nature.

[2] Price et al 2018. Bayesian Synthetic Likelihood. JCGS.

Bayesian Synthetic Likelihood

Basic method

- Simulate n iid datasets from the model based on θ
- Calculate the n sets of summary statistics
- Calculate the sample mean, μ_n, and sample covariance matrix, Σ_n, of the set of simulated summary statistics
- The BSL replacement likelihood is

$$\mathcal{N}(s_y; \mu_n(\theta), \Sigma_n(\theta)).$$

- Only tuning parameter is n (we find weak dependence on this choice). Choose n to maximise computational efficiency.

Asymptotic Properties of BSL[1]

Assuming summary stats follow a CLT (and other mild conditions):

- SL posterior mean is consistent and asymptotically normal.
- Asymptotically, BSL is more efficient than ABC.

[1] Nott et al 2019. Bayesian inference using synthetic likelihood: asymptotics and adjustments. arXiv:1902.04827

Drawbacks of BSL

- The distribution of the summary statistic must be roughly normal.
- The number of simulations per iteration, n, needs to be large depending on the size of the summary statistic to obtain a good sample covariance estimate.
- Reliance on MCMC to explore parameter space (not ideal in high dimensions).
- Computationally inefficient when model cannot recover observed statistic (e.g. model misspecification).

Shrinkage Covariance Estimation

When sample size n is small, the sample covariance matrix can have poor properties.

We propose to use shrinkage covariance matrix estimation to improve efficiency of BSL, e.g. the following estimator[1]

$$\hat{\Sigma}_\gamma = \hat{D}^{1/2}(\gamma\hat{C} + (1-\gamma)I)\hat{D}^{1/2}$$

where $\hat{C}$ is sample correlation matrix, $\hat{D}$ is the diagonal matrix with diagonal entries same as $\hat{\Sigma}$ and γ is the shrinkage parameter.

Amount of shrinkage trades-off the accuracy of the posterior distribution (relative to BSL) against computational efficiency.

[1] Warton 2008. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. JASA.

BSL with Whitening Transformation

Warton shrinkage is much more effective if summaries have weak correlation.

Assume $s \sim \mathcal{N}(0, \Sigma)$. We[1] consider the following transformation to summaries.

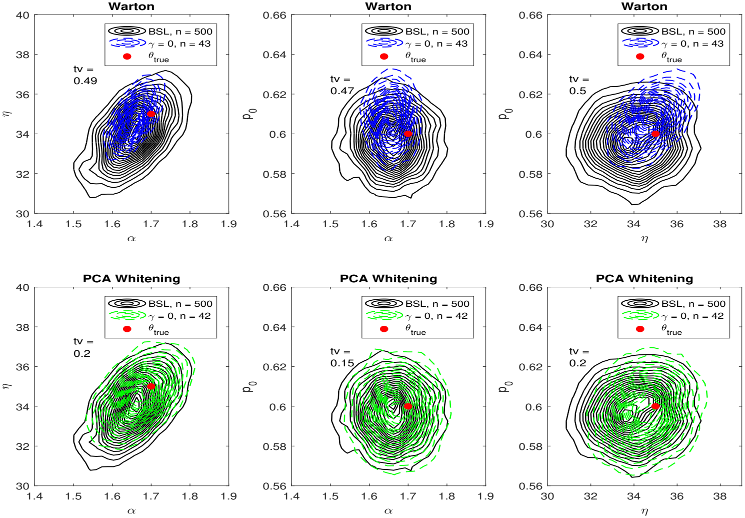
Let $\tilde{s} = Ws$ with $W = \Lambda^{-1/2} U^\top$ where Λ and U come from eigendecomposition of Σ (PCA whitening).

$\tilde{s}$ is then standard Gaussian. The Warton shrinkage is applied to $\tilde{s}$.

Σ^* is obtained by many off-line simulations from point estimate θ^*. We find that PCA whitening decorrelates well away from θ^*.

[1] Priddle et al 2019. Efficient Bayesian synthetic likelihood with whitening transformations. https://arxiv.org/abs/1909.04857.

Results - Toad Example (Model 1 Simulated Data)



Semi-parametric BSL

We improve flexibility of synthetic likelihood by using a semi-parametric Gaussian Copula model (called semiBSL[1]).

We use a kernel density estimate for the marginals. Capture dependence with Gaussian copula.

No additional tuning parameters and seemingly does not require any additional model simulations compared to BSL.

[1] An et al 2019. Robust Bayesian Synthetic Likelihood via a Semi-Parametric Approach. STCO.

Variational Bayes Synthetic Likelihood

Can significantly reduce number of model simulations replacing MCMC with Variational Bayes (VB).

Simulation problem $\rightarrow$ optimisation problem. E.g. finding the best $\mathcal{N}(\mu, \Sigma)$ representation of the posterior.

We have two papers[1][2] developing VB methods using synthetic likelihood.

Improves efficiency and better at dealing with high-dimensional parameter spaces. But resorts to Gaussian approximation of (approximate) posterior.

[1] Ong et al 2018. Variational Bayes with synthetic likelihood. STCO.

[2] Ong et al 2018. Likelihood-free inference in high dimensions with synthetic likelihood. CSDA.

BSL for Model Misspecification

BSL performs inefficiently when model is not compatible with observed statistic.

We are interested in modifications to the synthetic likelihood that allow BSL to run in such cases.

Also interested in identifying which statistics the model is not compatible with – might inform model refinement. [1]

[1] Frazier and Drovandi 2019. Robust Approximate Bayesian Inference with Synthetic Likelihood. https://arxiv.org/abs/1904.04551

Adjusted Synthetic Likelihoods

Consider two adjusted synthetic likelihoods for detecting incompatible summaries by introducing auxiliary variables γ of dimension $\mathrm{dim}(s)$.

1. Variance Inflation (R-BSL-V):

$$V_m(\zeta) := \Sigma_m + \begin{pmatrix} [\Sigma_m(\theta)]_{11}\gamma_1^2 & 0 & \cdots & 0 \\ 0 & [\Sigma_m(\theta)]_{22}\gamma_2^2 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & \cdots & [\Sigma_m(\theta)]_{d_\eta d_\eta}\gamma_{d_\eta}^2 \end{pmatrix},$$

2. Mean Adjustment (R-BSL-M):

$$\phi_m(\zeta) = \mu_m(\theta) + \mathrm{diag}(\Sigma_m^{1/2}(\theta))\Gamma,$$

where Γ is vector of γ's.

Theoretical Properties

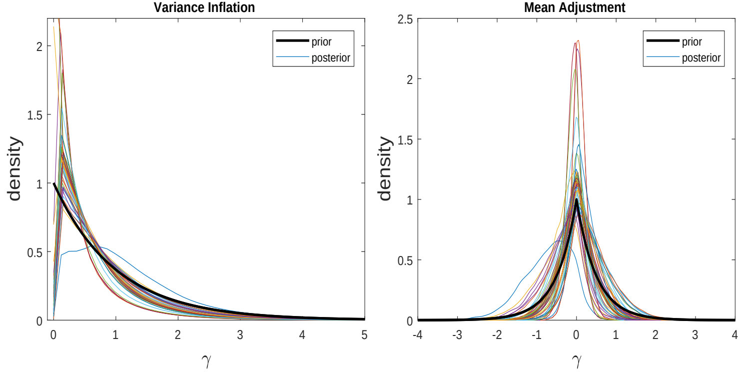
If the model is compatible, what behavior should we expect from the R-BSL approach?

Under compatibility and other mild assumptions, we show that the posterior for Γ converges to the prior.

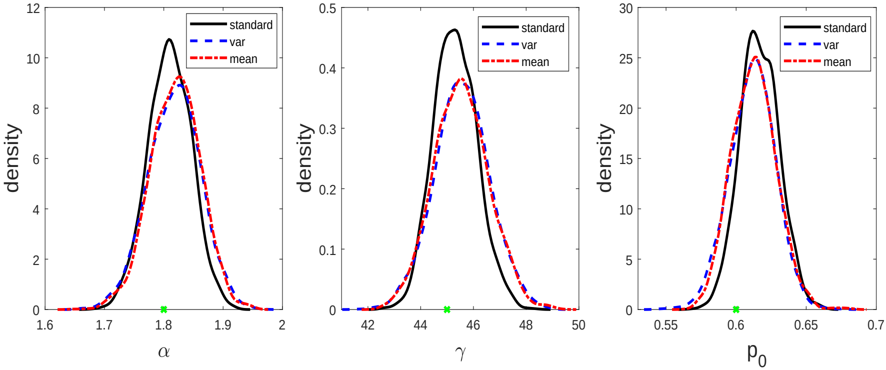
Thus incompatibility can be detected by departures from the prior.

Toad: Results for Simulated Data from Model 2

Consider model 2 as Marchand et al demonstrate that this model provides worse fit to data.
The MCMC acceptance rates for BSL, R-BSL-M and R-BSL-V are 11%, 9% and 22%, respectively ($n = 300$ for all methods).
Posteriors for γ:



Toad: Results for Simulated Data from Model 2

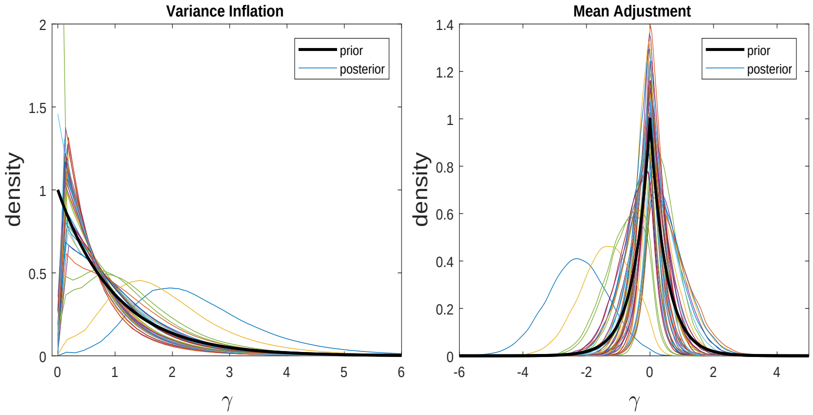


Univariate posterior distributions for the parameters when applying BSL (sold), R-BSL-V (dash) and R-BSL-M (dot-dash) to simulated data for the toad example. True parameter values are shown as crosses.

Toad: Results for Real Data (Model 2)

MCMC acceptance rates of 9% (BSL with $n = 2000$), 7% (R-BSL-M with $n = 500$) and 15% (R-BSL-V with $n = 500$). R-BSL-V has roughly 1 order of magnitude computational efficiency gain over BSL.

Posteriors for γ:



R Package

We have an evolving R package for BSL:

`https://cran.r-project.org/web/packages/BSL/index.html`

Associated paper:

`https://arxiv.org/abs/1907.10940`

Our Synthetic Likelihood Papers

Price et al (2018). Bayesian synthetic likelihood. JCGS.

An et al (2019). Accelerating Bayesian synthetic likelihood with the graphical lasso. JCGS.

An et al (2019). Robust Bayesian Synthetic Likelihood via a Semi-Parametric Approach. Statistics and Computing.

Priddle et al 2019. Efficient Bayesian synthetic likelihood with whitening transformations. https://arxiv.org/abs/1909.04857.

Nott et al (2019). Bayesian inference using synthetic likelihood: asymptotics and adjustments. https://arxiv.org/abs/1902.04827

Frazier and Drovandi (2019). Robust Approximate Bayesian Inference with Synthetic Likelihood. https://arxiv.org/abs/1904.04551

Ong et al (2018). Variational Bayes with Synthetic Likelihood. Statistics and Computing.

Ong et al (2018). Likelihood-Free Inference in High Dimensions with Synthetic Likelihood. CSDA.

An et al (2019). BSL: An R Package for Efficient Parameter Estimation for Simulation-Based Models via Bayesian Synthetic Likelihood. https://arxiv.org/abs/1907.10940

Email: c.drovandi@qut.edu.au Web: chrisdrovandi.weebly.com Twitter: @chris_drovandi