# Bayesian Synthetic Likelihood

Thanasi Bakis, Brian Schetzsle

March 18, 2022

## Overview

At a high level, the synthetic likelihood function is a replacement for an intractable true likelihood. Synthetic likelihood originates in a frequentist setting (Wood 2010), when the likelihood function is too irregular to easily maximize (analytically or numerically) to obtain a point estimator for parameters. Assuming we can at least generate samples from the true likelihood, we can maximize the synthetic likelihood instead, which yields an alternative point estimator for the parameters. It is trivial to extend this idea to a Bayesian setup by placing a prior on the parameters, obtaining a posterior distribution of parameters instead of point estimates (Price et al. 2018). We explore this idea of a Bayesian synthetic likelihood method.

## Motivation

Both Wood (2010) and Price et al. (2018) include the Ricker population model as a motivating example for the frequentist and Bayesian synthetic likelihood methods, respectively. However, while the former illustrates the difficulty of maximizing the likelihood in the Ricker model, the latter focuses more on applying the method without thoroughly addressing the concerns using standard MCMC with this model. As a result, we attempt to seek an understanding ourselves.

In the Ricker model, $N_t$ models the time course of a population, where $N_0 = 1$ and $N_t | N_{t-1} \sim LogNormal(\log N_{t-1} + \log r - N_{t-1}, \sigma^2)$. Additionally, suppose $N_t$ is not actually observed, but a sample from the population $Y_t | N_t \sim Poisson(\phi N_t)$ is observed. Our parameters are thus: $r$, a population growth rate parameter; $\sigma$, controlling random noise; and $\phi$, a scaling parameter for sampling from the population.

To do Bayesian inference via Metropolis-Hastings, we need to compute the posterior, up to the normalizing constant: $p(r, \sigma, \phi | \boldsymbol{Y}) \propto p(r, \sigma, \phi) p(\boldsymbol{Y} | r, \sigma, \phi)$. The question remains whether we can obtain the likelihood $p(\boldsymbol{Y} | r, \sigma, \phi)$ required.

Let $\boldsymbol{\theta} = (r, \sigma, \phi)$ for notational simplicity. While the marginal likelihood with the observed $\boldsymbol{Y}$ is not obvious, the joint likelihood with observed and latent variables is easier to work out:

$$\begin{aligned} &p(\boldsymbol{Y}, \boldsymbol{N} | \boldsymbol{\theta}) \\ &\quad = p(\boldsymbol{N} | \boldsymbol{\theta}) p(\boldsymbol{Y} | \boldsymbol{N}, \boldsymbol{\theta}) \\ &\quad = p(N_1 | \boldsymbol{\theta}) \prod_{t=2}^{n} p(N_t | N_{t-1}, \boldsymbol{\theta}) \prod_{t=1}^{n} p(Y_t | N_t, \boldsymbol{\theta}) \end{aligned}$$

which is a product of log-normal densities and Poisson mass functions. This cannot be evaluated, though, since it requires values of $N_t$ that are unobserved. To obtain the marginal likelihood without unobserved variables, we would need to marginalize over all $N_t$:

$$p(\boldsymbol{Y} | \boldsymbol{\theta}) = \int_{N_1} \ldots \int_{N_n} p(\boldsymbol{Y}, \boldsymbol{N} | \boldsymbol{\theta}) \, dN_1 \ldots dN_n$$

This is expensive, in part because the number of integrals grows with the number of data points observed.

Alternatively, instead of having the posterior $p(r, \sigma, \phi | \boldsymbol{Y})$, we could also try to do inference on the latent population variables: $p(N_1, \ldots, N_n, r, \sigma, \phi | \boldsymbol{Y})$. However, this would require the dimension of Metropolis-Hastings proposals to grow with the number of observed data points too (proposing values of each $N_t$). We think this is undesirable; Price et al. (2018) does not address the drawbacks of traditional MCMC for this model over using the synthetic likelihood.

## Methodology

We now outline the method Price et al. (2018) refers to as “Markov chain Monte Carlo Bayesian synthetic likelihood” (MCMC-BSL), which we refer to as BSL.

BSL is essentially a Metropolis-Hastings algorithm with a different target distribution than the traditional posterior. Instead of targeting the posterior $p(\boldsymbol{\theta}|\boldsymbol{Y})$ in our MCMC, BSL targets $p(\boldsymbol{\theta}|\boldsymbol{s_Y})$, where $\boldsymbol{s_Y}$ is a vector of summary statistics for $\boldsymbol{Y}$ (eg. the mean, quantiles, etc.). This would use a so-called "synthetic likelihood" $p(\boldsymbol{s_Y}|\boldsymbol{\theta})$ in place of the true likelihood $p(\boldsymbol{Y}|\boldsymbol{\theta})$ when applying Bayes' rule, allowing us to circumvent an intractable true likelihood.

It is important to note that the new target posterior does not necessarily approximate the original; it is more of an alternative choice. This is similar to obtaining a maximum synthetic likelihood estimator instead of an MLE in the frequentist setting. The MSLE doesn't necessarily approximate the MLE; it is an alternative.

Of course, depending on your choice of statistics, this synthetic likelihood may not be tractable either. Consequently, BSL makes a normality assumption: $\boldsymbol{s_Y}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ change with $\theta$. Ideally, the choice of statistics truly are normally distributed, but this may not be the case. We will see examples of both cases later.

While the normality assumption presents us with a tractable distribution, we still cannot evaluate this likelihood because of the unknown $\boldsymbol{\mu_\theta}$ and $\boldsymbol{\Sigma_\theta}$ parameters. BSL outlines a method for estimating these parameters for any proposed value of $\theta$: *under the assumption that we can sample $\boldsymbol{Y}|\boldsymbol{\theta}$ from the true original likelihood*, the parameters can be estimated via Monte Carlo approximation. Let $\boldsymbol{Y_1^*}, ..., \boldsymbol{Y_n^*} \overset{\text{iid}}{\sim} p(\boldsymbol{Y}|\boldsymbol{\theta})$. In other words, given a proposed value of $\boldsymbol{\theta}$, generate $n$ iid datasets from the same family of distributions that also generated the observed data $\boldsymbol{Y}$. For each simulated dataset $\boldsymbol{Y_i^*}$, calculate the corresponding summary statistics $\boldsymbol{s_{Y_i^*}}$ in the same manner as $\boldsymbol{s_Y}$. Then, estimate the synthetic likelihood parameters as:

$$\hat{\boldsymbol{\mu}_{\boldsymbol{\theta}}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s_{Y_i^*}}$$

$$\hat{\boldsymbol{\Sigma}_{\boldsymbol{\theta}}} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{s_{Y_i^*}} - \hat{\boldsymbol{\mu}_{\boldsymbol{\theta}}})(\boldsymbol{s_{Y_i^*}} - \hat{\boldsymbol{\mu}_{\boldsymbol{\theta}}})^\top$$

With these estimates, we can evaluate the synthetic likelihood of various proposed values of $\boldsymbol{\theta}$ for our observed summary statistics $\boldsymbol{s_Y}$, which, combined with a prior over $\boldsymbol{\theta}$, completes the requirements for performing Bayesian inference via Metropolis-Hastings.

## Experiments

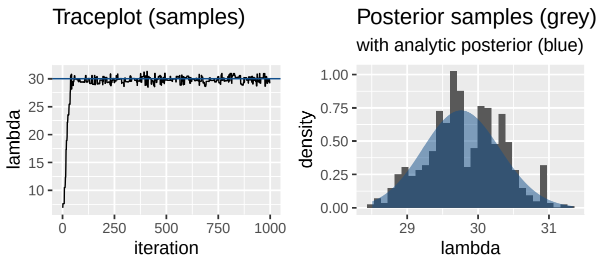
Price et al. (2018) first elaborates on the mechanism of BSL using a toy example involving a gamma-Poisson Bayesian model. They then demonstrate the method with the Ricker model introduced in the motivating example earlier. After implementing the algorithm ourselves, we attempted to reproduce these examples, as well as extend the toy example in an insightful manner.

### Gamma-Poisson model

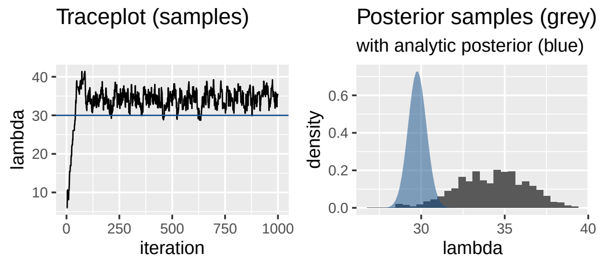
Consider data $\boldsymbol{Y}$ drawn from a Poisson distribution, $Poisson(\lambda)$ and a prior distribution placed on $\lambda \sim Gamma(\alpha, \beta)$. We are interested in performing posterior inference on $p(\lambda|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\lambda)p(\lambda)$. Suppose we observe a sample of size $N = 100$ generated using $\lambda = 30$, $\alpha = \beta = 0.001$. It is important to note that the true posterior can be derived analytically as $\lambda|\boldsymbol{Y} \sim Gamma(\alpha = 0.001 + \sum_{i=1}^{100} Y_i, \beta = 100.001)$. Thus, for any choice of statistic $\boldsymbol{s_Y}$, we can compare the approximated distribution of $\lambda|\boldsymbol{s_Y}$ from BSL to the known, true posterior conditional on the full data, $\lambda|\boldsymbol{Y}$.

The authors acknowledge that this is not always the aim of BSL. However, when $\boldsymbol{s_Y}$ is a sufficient statistic for the sample $\boldsymbol{Y}$, the two distributions are the same. Price et al. (2018) chooses the statistic to be the sample mean, $s_Y = \overline{Y}$, which is sufficient for Poisson samples. The sample mean also has the added benefit of approximate normality guaranteed by the central limit theorem, so the normality assumption behind BSL is satisfied.

We run BSL and generate 1000 posterior samples of $\lambda$. We start at $\lambda = 6$, far from the actual $\lambda = 30$, so that we can see BSL converging to the true value. The resulting traceplot, as well as a histogram of posterior samples compared to the true analytic posterior, are shown below.

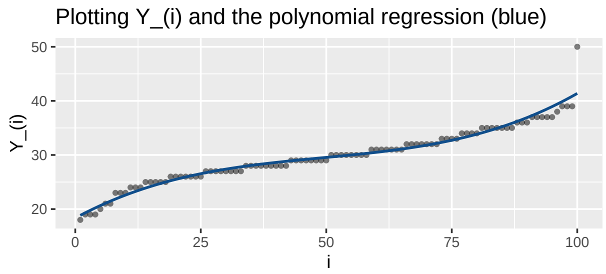


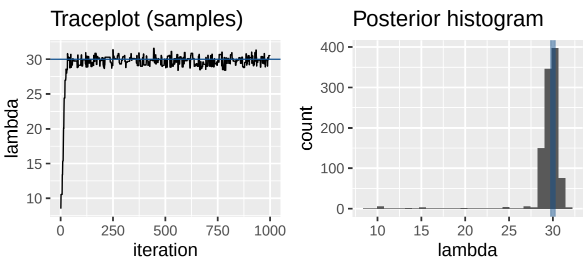
With a functioning implementation of BSL, it is trivial to swap in different summary statistics. We continued to explore this toy example beyond what was presented by Price et al. (2018), next selecting the maximal value for our statistic, $s_Y = \max(\boldsymbol{Y})$. This statistic is appealing because it is easy to compute, but is neither sufficient nor normally distributed, allowing us to explore the impact of a poorly-selected statistic on posterior inference. Results of the simulation are below.



This choice of statistic resulted in posterior samples of $\lambda | s_Y$ that are radically different from the distribution of $\lambda | \boldsymbol{Y}$. As it turns out, the maximum value in our data, $\boldsymbol{Y_{(100)}}$, was unusually large, so the MCMC algorithm was identifying larger values of $\lambda$ as more likely to have generated such a large maximal statistic than the true value of $\lambda = 30$. This reified for us the sensitivity of BSL on the choice of statistic. Further, once a statistic is chosen, if the observed data has an unusual value for this statistic, posterior inference can radically differ from the truth.

We finally tried to choose a statistic that could capture much of the nuance in our model without being sufficient. We settled on the coefficients of a polynomial regression of the ordered observations: $Y_{(i)} = \beta_0 + \beta_1 i + \beta_2 i^2 + \beta_3 i^3$. The idea is, by fitting a curve to the ordered data, we can summarise the distribution of the data using the coefficients of the fit. Below is a plot of the ordered observations with the fitted regression line (the unusual value $Y_{(100)} = 50$ is clearly visible). This curve indeed captures the general shape of the data, and our simulation confirms that this choice of statistic does lead to posterior samples of $\lambda | \boldsymbol{s_Y}$ that more closely resemble $\lambda | \boldsymbol{Y}$.





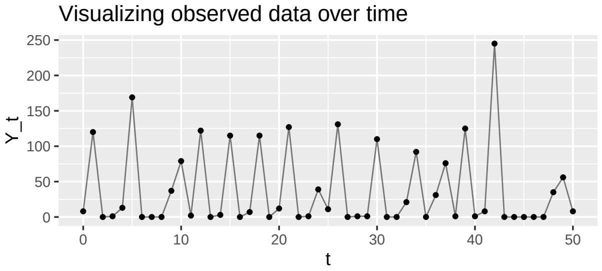
## Ricker population model example

We now return to the Ricker model from the original motivating example to provide a more complex use case. In contrast to the toy Poisson model, the sufficient statistics of the Ricker model are not easily available. Furthermore, we will need to rely on more complex, non-normal statistics to summarise the data.

Suppose we observe $Y_1, ..., Y_{50}$ with $(\log r, \sigma, \phi) = (3.8, 0.3, 10)$. We place an uninformative prior over all parameters.

As mentioned earlier, to use BSL, we need to ensure we can sample simulated $\boldsymbol{Y}|\boldsymbol{\theta}$ from the model. Thankfully, sampling is straightforward; using $N_0 = 1$, you can sample $N_1, N_2, ...$ sequentially, then sample $Y_t$ for each $N_t$.

We visualize the observed data $\boldsymbol{Y}$ below:



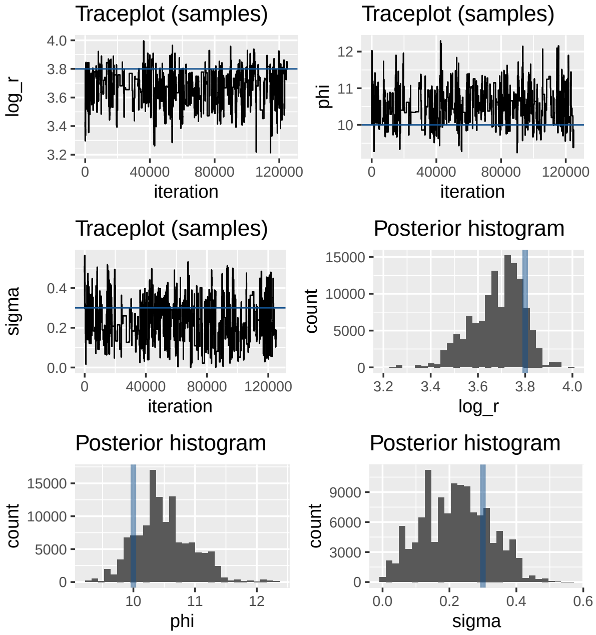
This data is noisy and runs over time, yielding the question: what sorts of statistics will we use? Wood (2010) proposes numerous statistics that can be classified into three groups, and Price et al. (2018) follows suit, as do we.

**Marginal distribution statistics** These summarize the "shape" of the marginal distribution. We use two of these statistics: the sample mean, $\overline{Y}$; and the number of zeroes observed, $\sum_{t=1}^{50} \mathbf{1}_{\{0\}}(Y_t)$.

**Dynamic process statistics** These characterize the relationship between $Y_t$ and $Y_{t-1}$ (and possibly more history). We use the coefficients of two regression models as statistics: an autoregressive model, $Y_t^{0.3} \sim Y_{t-1}^{0.3} + Y_{t-1}^{0.6}$; and a regression on ordered differences, $(Y_t - Y_{t-1}) \sim Y_t + Y_t^2 + Y_t^3$. Note that the exponents of the autoregressive model were tuned by Wood (2010) to improve fit.

**Time series statistics** These are sensitive to the shape and period of fluctuations. We use the coefficients of the autocovariance function, up to lag 5, as statistics.

The results of BSL on this model with all these statistics together are showen below:



As we can see, the chains appear to converge, but some of the distributions are a bit shifted from the true parameter values. This could be due to our data having extreme statistic values, a phenomenon seen earlier in the Poisson toy example with the maximal statistic. It is also worth noting that we ran BSL on this model for around 125,000 iterations, whereas Price et al. (2018) ran it for 500,000 iterations.

# Discussion

## Alternative methodology

A common method used when likelihoods are intractable is approximate Bayesian computation (ABC) (Sisson 2011). ABC is very similar to Bayesian synthetic likelihood (BSL), but it replaces the normality assumption of $\boldsymbol{s_Y}|\boldsymbol{\theta}$ with a nonparametric likelihood:

$$p(\boldsymbol{s_Y}|\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} K_\epsilon(\rho(\boldsymbol{s_Y}, \boldsymbol{s_{Y_i^*}}))$$

where $\boldsymbol{Y_i^*}$ are the Monte Carlo simulated datasets, $\rho$ is a distance measure, and $K_\epsilon$ is a kernel weighting function with bandwidth $\epsilon$.

Price et al. (2018) offers a few points of comparison between BSL and ABC: (1) the BSL posterior fell in the vicinity of the ABC posterior for the Ricker example, despite the non-normality of the summary statistics used; (2) ABC has two tuning parameters (the number of Monte Carlo samples $n$, and the bandwidth $\epsilon$), where as BSL only has one ($n$); and (3) the curse of dimensionality impacts ABC more than BSL due to its nonparametric nature.

Note that we did not implement ABC, since the Price et al. (2018) noted that experiments involving ABC needed to be run for 25 million iterations.

## Reflecting on this paper

Finally, we briefly discuss some strengths and points of improvements for Price et al. (2018).

Regarding strengths of the paper, we appreciated that the authors provided implementations for their algorithm and experiments. While we found their codebase after implementing the algorithm and examples ourselves, it proved instrumental in debugging our code and getting the examples running. Additionally, we applauded the examples provided in the paper, which appropriately increased in complexity to demonstrate BSL (from having an analytical posterior available, to lacking this and requiring a violation of the normality assumption).

For points of improvement, we feel the authors did not seem to thorougly discuss the limitations of BSL, particularly when it came to the choice of statistic. As previously discussed, we did a bit of further exploration with the Poisson example to see which choices of statistics could cause BSL to underperform. Additionally, we feel the authors did not extensively motivate the use of a synthetic likelihood in a Bayesian setting very well. While Wood (2010) explained the difficulties using the maximum likelihood framework for the Ricker example, Price et al. (2018) seemed to merely extend the synthetic likelihood framework to the Bayesian setting without thorough reason.

## References

Price, Leah F, Christopher C Drovandi, Anthony Lee, and David J Nott. 2018. "Bayesian Synthetic Likelihood." *Journal of Computational and Graphical Statistics* 27 (1): 1–11.

Sisson, Scott A. 2011. "Likelihood-Free Markov Chain Monte Carlo." *Handbook of Markov Chain Monte Carlo*.

Wood, Simon N. 2010. "Statistical Inference for Noisy Nonlinear Ecological Dynamic Systems." *Nature* 466 (7310): 1102–4.