Bayesian Synthetic Likelihood

Thanasi Bakis, Brian Schetzsle

March 17, 2022

Original paper(s)

- Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature.*
  - Introduced *synthetic likelihood* technique in a frequentist setting
- Price, L. F., Drovandi, C. C., Lee, A., & Nott, D. J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics.*
  - Extended Wood (2010) to *Bayesian synthetic likelihood* technique in a Bayesian setting

Motivating the synthetic likelihood

At a high level, the synthetic likelihood is a replacement for an intractable true likelihood.

Synthetic likelihood originates in a *frequentist setting*, when the likelihood function is too irregular to easily maximize (analytically or numerically).

Maximizing the synthetic likelihood yields a point estimator for your parameters.

Motivating the synthetic likelihood

It is trivial to extend this to a Bayesian setup by placing a prior on the parameters, obtaining a posterior distribution of parameters instead of point estimates.

The core requirement of the synthetic likelihood method is that we can still generate samples from the true likelihood.

Upcoming motivating example from the paper: Ricker population model.

(Note that the paper does not thoroughly address the concerns with this motivating example. We attempt to seek an understanding ourselves.)

Motivating the synthetic likelihood

In the Ricker model, N_t models the time course of a population, where

$$\begin{aligned} N_0 &= 1 \\ N_t &= rN_{t-1}\mathrm{e}^{-N_{t-1}+e_{t-1}} \\ e_t &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \end{aligned}$$

In other words, $N_t|N_{t-1} \sim LogNormal(\log N_{t-1} + \log r - N_{t-1}, \sigma^2)$.

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; σ, controlling random noise; and ϕ, a scaling parameter for sampling from the population.

Motivating the synthetic likelihood

To do Bayesian inference, we need:

$$p(r, \sigma, \phi|\boldsymbol{Y}) \propto p(r, \sigma, \phi)p(\boldsymbol{Y}|r, \sigma, \phi)$$

Can we obtain the likelihood $p(\boldsymbol{Y}|r, \sigma, \phi)$ required?

Motivating the synthetic likelihood

Let $\boldsymbol{\theta} = (r, \sigma, \phi)$. The joint likelihood with observed and latent variables is easier to work out:

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

Motivating the synthetic likelihood

We would need to marginalize over all N_t:

$$p(\boldsymbol{Y}|r, \sigma, \phi) = \int_{N_1} ... \int_{N_n} p(\boldsymbol{Y}, \boldsymbol{N}|\boldsymbol{\theta}) \, dN_1 ... dN_n$$

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

Motivating the synthetic likelihood

Alternatively, instead of having the posterior $p(r, \sigma, \phi|\boldsymbol{Y})$, we could also try to do inference on the latent population variables: $p(N_1, ..., N_n, r, \sigma, \phi|\boldsymbol{Y})$.

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; the paper does not address the drawbacks of traditional MCMC for this model over using the synthetic likelihood.)

Bayesian synthetic likelihood

What if, instead of targeting the posterior $p(\boldsymbol{\theta}|\boldsymbol{Y})$ in our MCMC, we target:

$$p(\boldsymbol{\theta}|\boldsymbol{s_Y})$$

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 $p(\boldsymbol{Y}|\boldsymbol{\theta})$.

(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.)

Bayesian synthetic likelihood

Of course, depending on your choice of statistics, this *synthetic likelihood* may not be tractable either. This method consequently 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 θ.

Ideally, your statistics truly are normally distributed, but this may not be the case. We will see examples of both cases.

Bayesian synthetic likelihood

$$\boldsymbol{s_Y}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu_\theta}, \boldsymbol{\Sigma_\theta})$$

We still need to be able to evaluate this likelihood, though.

Under the assumption that we can sample $\boldsymbol{Y}$ *from the real likelihood*, we can estimate the parameters of the synthetic likelihood via Monte Carlo approximation…

Suppose we have a proposed value of $\boldsymbol{\theta}$. 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 $\boldsymbol{Y}$.

Bayesian synthetic likelihood

Now, for each simulated dataset $\boldsymbol{Y_i^*}$, calculate the corresponding summary statistics $\boldsymbol{s_{Y_i^*}}$ in the same manner as $\boldsymbol{s_Y}$.

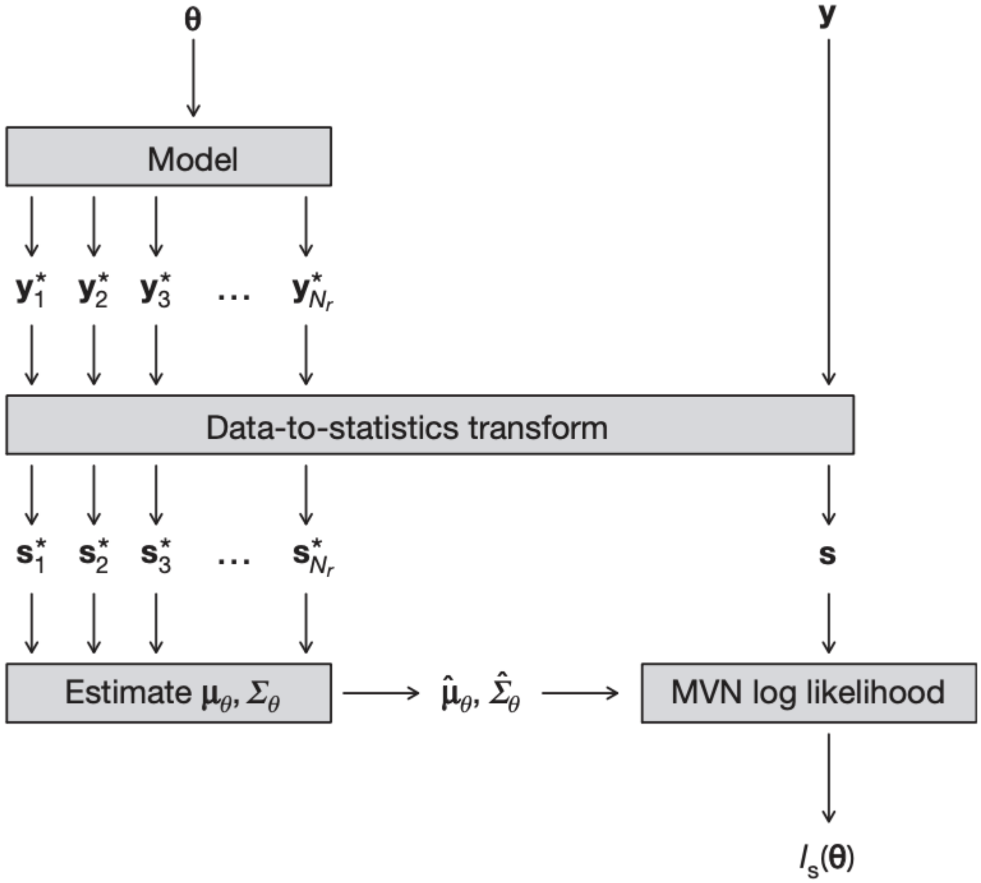
Then,

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

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

With these estimates, we can evaluate the synthetic likelihood of various proposed values of $\boldsymbol{\theta}$ for our observed summary statistics $\boldsymbol{s_Y}$.

Bayesian synthetic likelihood


θ
y
Model
y^*_1 y^*_2 y^*_3 … $y^*_{N_r}$
Data-to-statistics transform
s^*_1 s^*_2 s^*_3 … $s^*_{N_r}$
s
Estimate $\mu_\theta, \Sigma_\theta$
$\hat{\mu}_\theta, \hat{\Sigma}_\theta$
MVN log likelihood
$l_s(\theta)$

A Toy Example

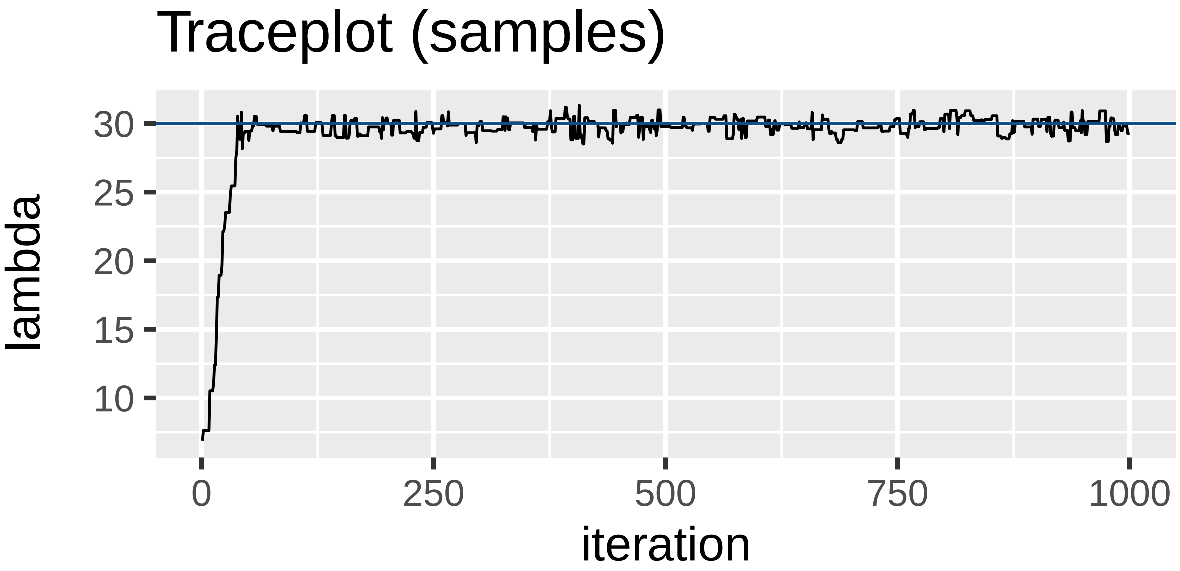
Price et. al. first elaborate on the mechanism of synthetic likelihood using an elementary toy example. Consider data drawn from a Poisson distribution, $Poisson(\lambda)$, with mean λ and a prior distribution placed on λ of $Gamma(\alpha, \beta)$. We are interested in performing posterior inference on $p(\lambda|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\lambda)p(\lambda)$ which is proportional to the conditional likelihood of the data, $p(\boldsymbol{Y}|\lambda)$, multiplied by the prior density of the parameter, $p(\lambda)$. The authors chose values $\lambda = 30$, $\alpha = \beta = 0.001$, and sample size of $N = 100$ which yields a posterior distribution that can adequately be approximated by the normal distribution assumed by BSL.

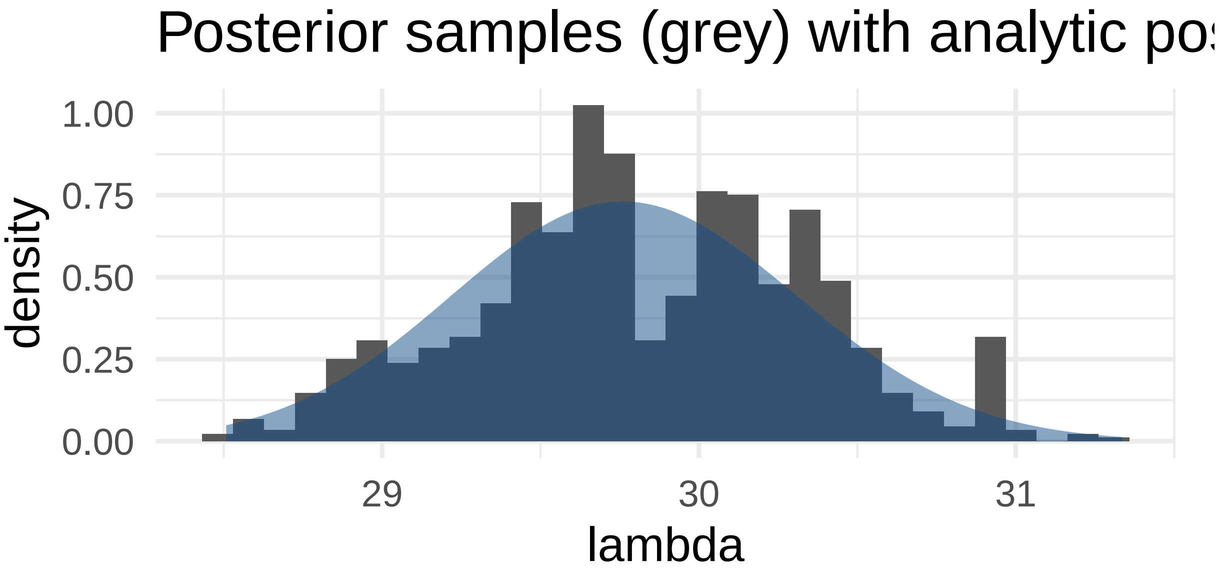
It is important to note that the true posterior can be derived analytically:

$$
\begin{aligned}
p(\lambda|\boldsymbol{Y}) &\propto \left[\prod_{i=1}^{100} \frac{\lambda^{Y_i}}{Y_i!} e^{-\lambda}\right] \left[\frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}\right] \\
&\propto \lambda^{\alpha+\sum_{i=1}^{100} Y_i - 1} e^{-\lambda(\beta+n)} \\
&\sim Gamma(\alpha = 0.001 + \sum_{i=1}^{100} Y_i, \beta = 100.001)
\end{aligned}
$$

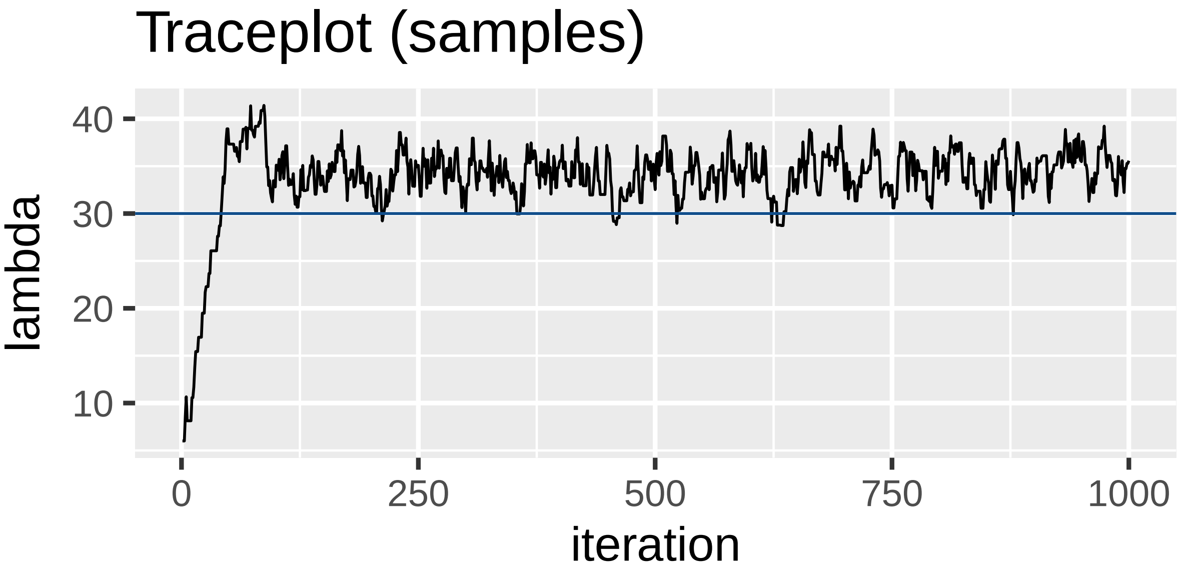
Thus, for any choice of statistic $\boldsymbol{s_Y}$ we can compare the distribution of posterior samples of $\lambda|\boldsymbol{s_Y}$ to the known, true posterior conditional on the full data, $\lambda|\boldsymbol{Y}$. The authors acknowledge that this is not always the aim of BSL. However, when $\boldsymbol{s_Y}$ is a sufficient statistic for the sample $\boldsymbol{Y}$ the two distributions are the same. Price et. al. use the mean, $\boldsymbol{s_Y} = \bar{Y}$, as the statistic which is sufficient for Poisson samples. The mean also has the added benefit that the central limit theorem guarantees that means of repeated samples are normally distributed so the central assumption behing BSL is satisfied automatically.

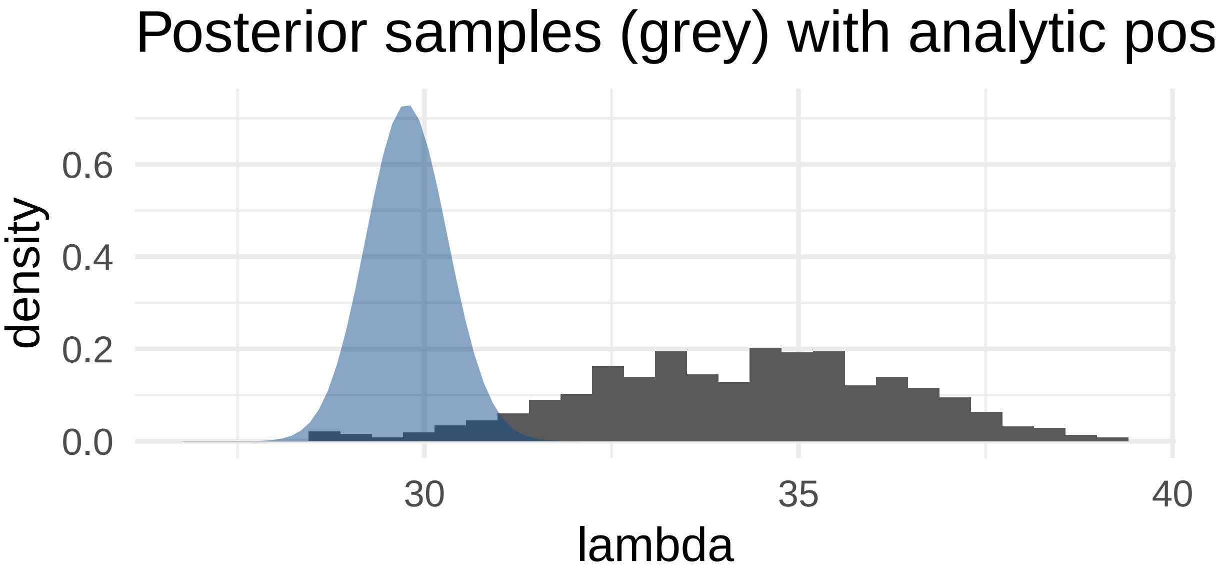
We replicate this setup and generate 1000 posterior samples of λ. We start at $\lambda = 6$, far from the actual $\lambda = 30$, so that we can see BSL converging to the true value. The resulting trace plot and histogram of posterior samples compared to the true analytic posterior are shown below.





Once we had our BSL code working it was trivial to swap in different summary statistics. We continued to explore this toy problem beyond what was presented by Price et. al., next selecting the maximal value for our statistic, $\boldsymbol{s_Y} = \max(\boldsymbol{Y})$. This statistic was 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.


Traceplot (samples)

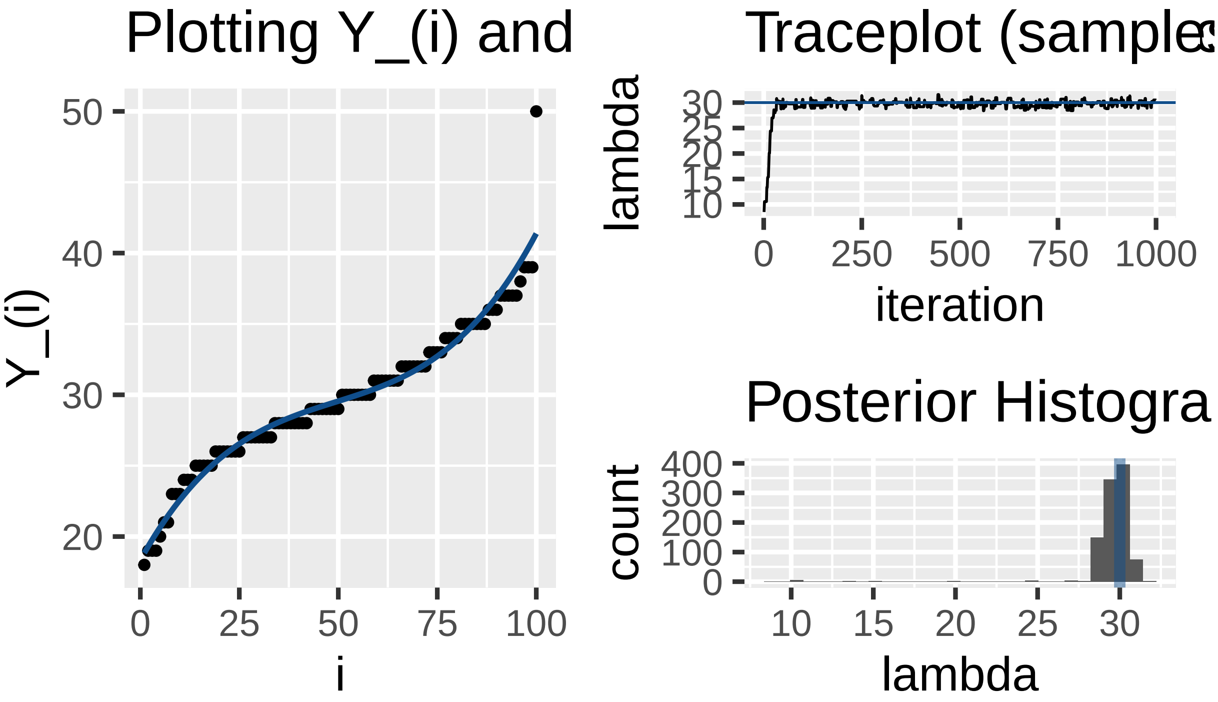

Posterior samples (grey) with analytic pos

The BSL substitution resulted in posterior samples of λ conditional on $\boldsymbol{s_Y}$ that are radically different from the distribution of $\lambda|\boldsymbol{Y}$. We looked into why this is the case and found that the maximum value in our data, $\boldsymbol{Y_{(100)}}$, was unusually large so the MCMC algorithm was identifying larger values of λ 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 this method 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

$$Y_{(i)} = \beta_0 + \beta_1 i + \beta_2 i^2 + \beta_3 i^3$$

Below is a plot of the ordered observations with the fitted regression line (the unusual value $Y_{(100)} = 50$ is clearly visible). This curve captures the general shape of the data and our simulation confirmed 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 beginning of the presentation to provide a more complex use case.

Some things to consider about the Ricker model vs the toy Poisson model:

- The sufficient statistics are not easily available
- We will need to rely on more complex, non-normal statistics

Ricker population model example

As a refresher, recall we observe a sample Y_t of a population N_t:

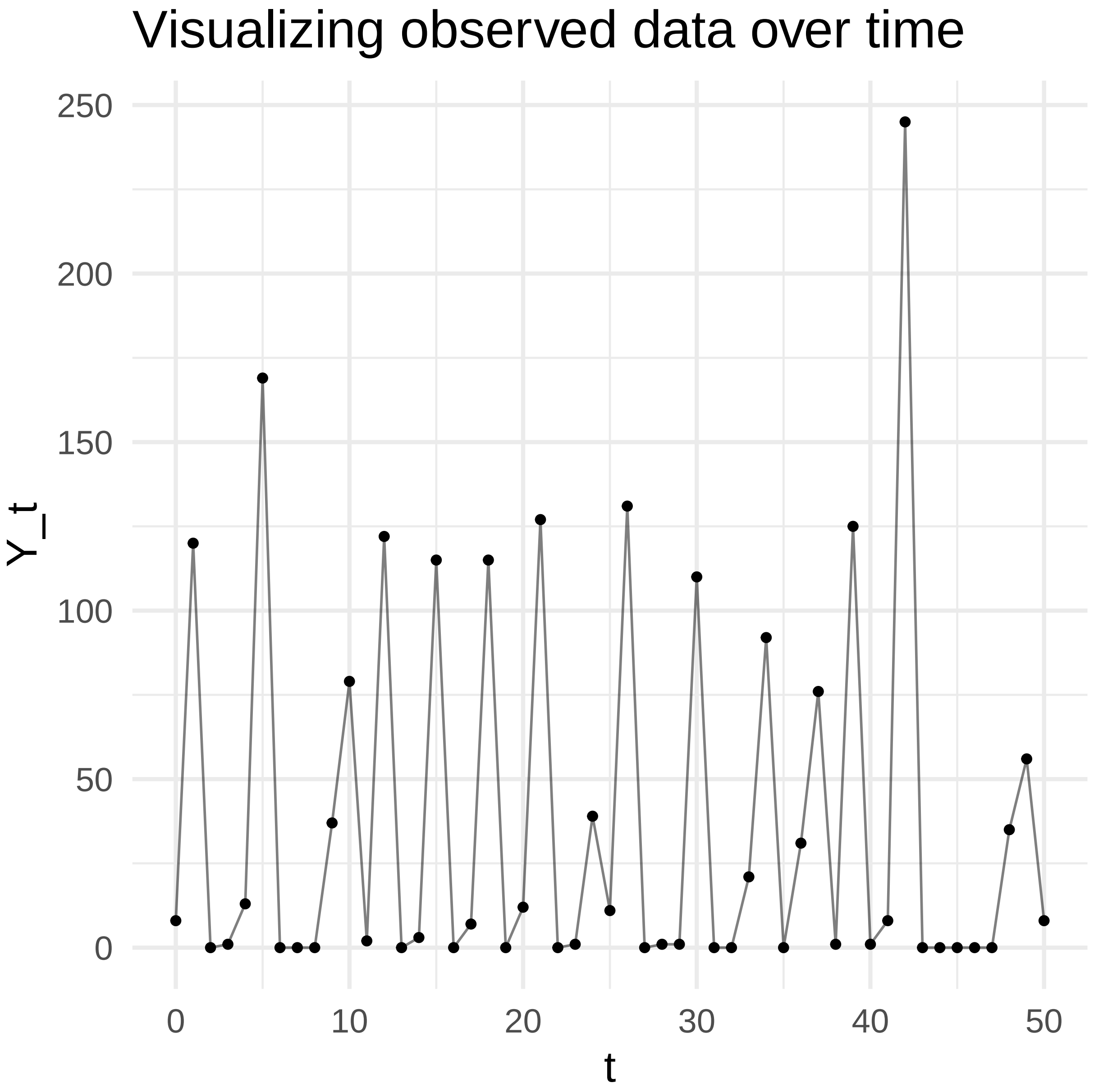
$$\begin{aligned} N_t|N_{t-1} &\sim LogNormal(\log N_{t-1} + \log r - N_{t-1}, \sigma^2) \\ Y_t|N_t &\sim Poisson(\phi N_t) \end{aligned}$$

Suppose we observe $Y_1, ..., Y_{50}$ with $(\log r, \sigma, \phi) = (3.8, 0.3, 10)$.

To use synthetic likelihood, we need to ensure we can sample simulated $\boldsymbol{Y}$ 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 place an uninformative prior over all parameters.

Ricker population model example



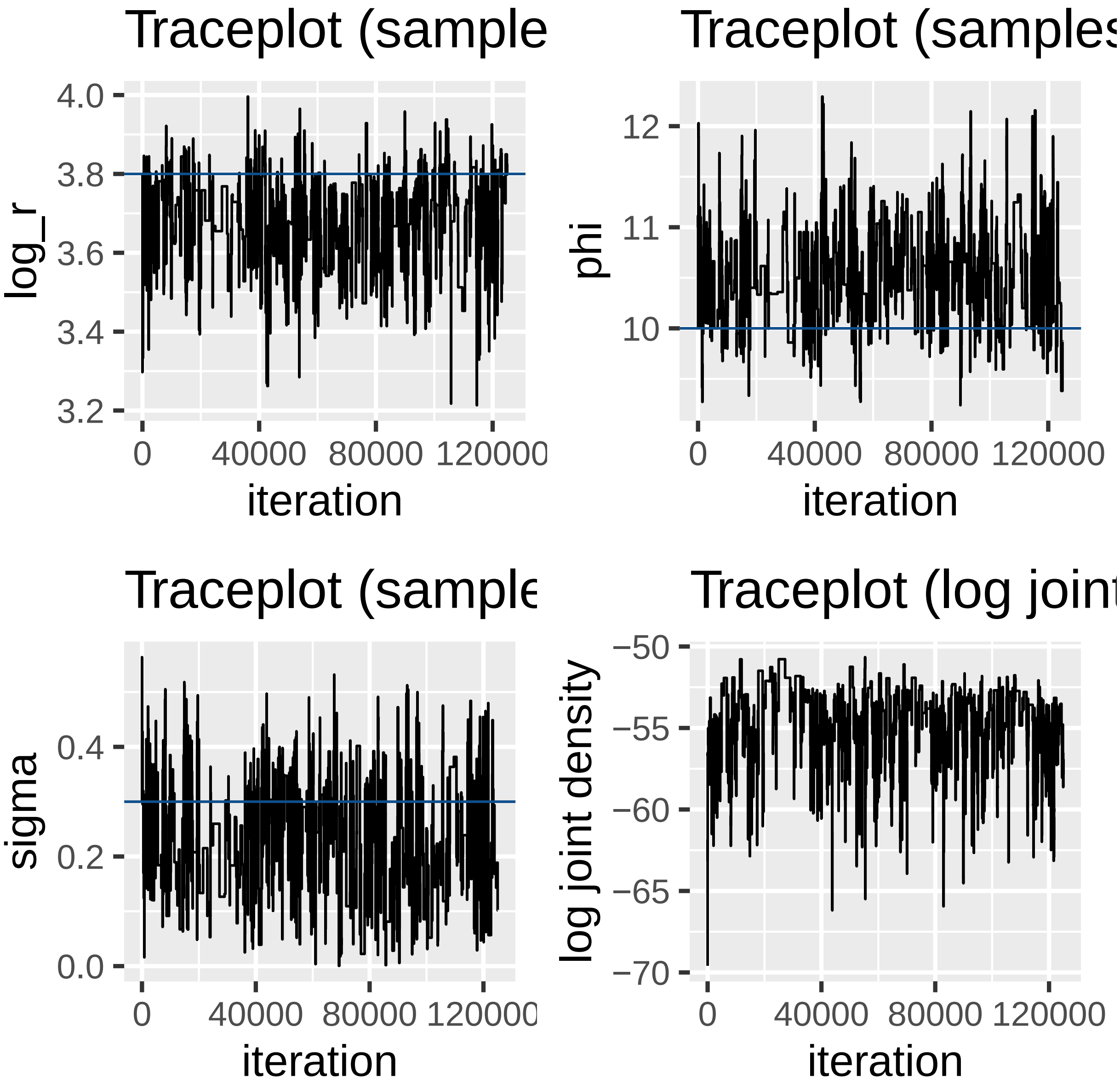
Ricker population model example

This data is noisy and runs over time. What sorts of statistics will we use?

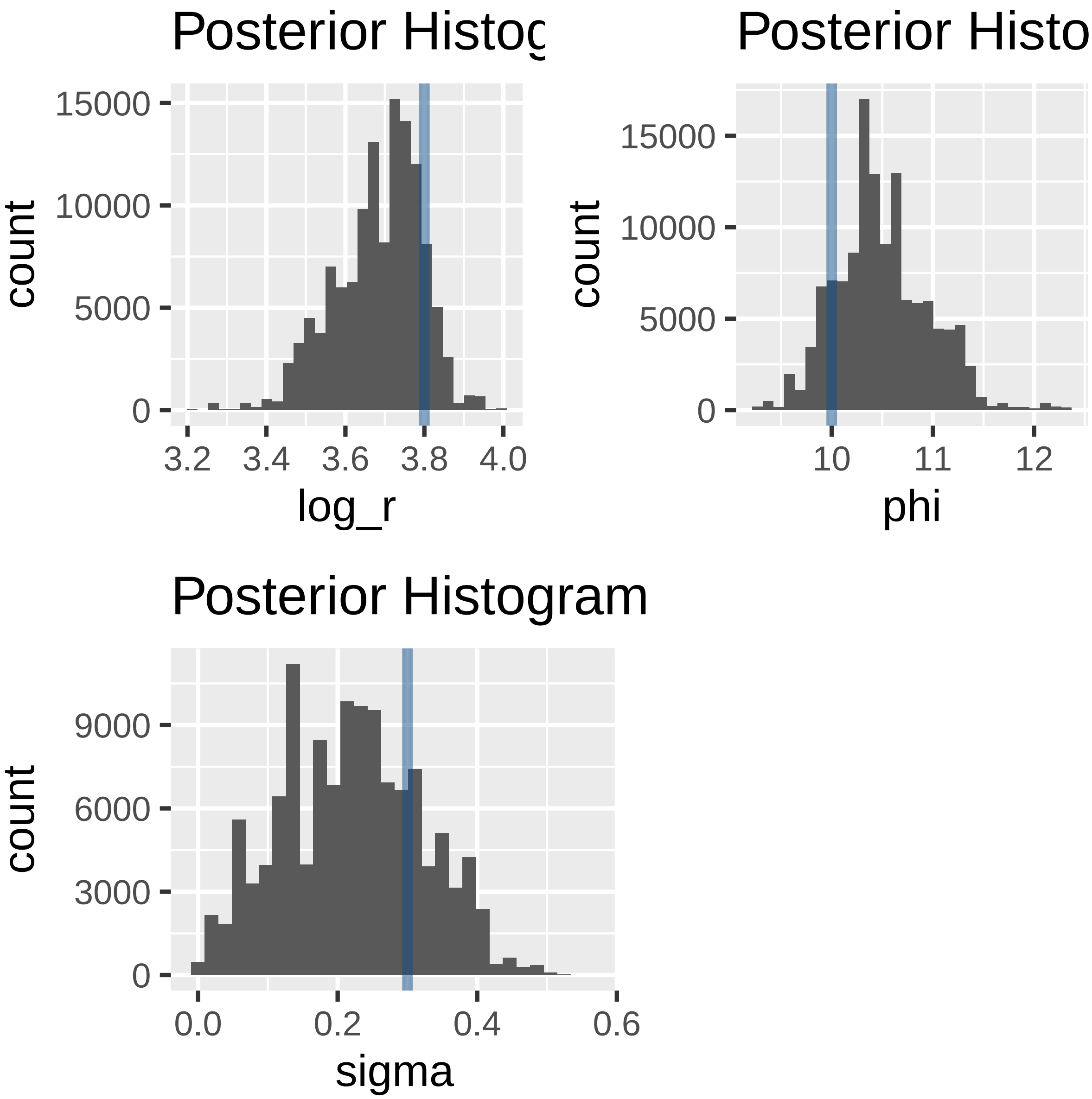
- **Marginal distribution statistics**: to summarize the "shape" of the marginal distribution
  - Mean: $\overline{Y}$
  - Number of zeros: $\sum_{t=1}^{50} \mathbf{1}_{\{0\}}(Y_t)$
- **Dynamic process statistics**: characterize the relationship between Y_t and Y_{t-1} (and possibly more history)
  - Autoregressive model coefficients: $Y_t^{0.3} \sim Y_{t-1}^{0.3} + Y_{t-1}^{0.6}$
    - Note the exponents were tuned to improve fit
  - Coefficients of regression on ordered differences: $(Y_t - Y_{t-1}) \sim Y_t + Y_t^2 + Y_t^3$

- **Time series statistics**: sensitive to the shape and period of fluctuations
    - Coefficients of the autocovariance function, up to lag 5

Ricker population model example



Ricker population model example



Alternative methodology

A common method used when likelihoods are intractable is approximate Bayesian computation (ABC).

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, ρ is a distance measure, and K_ϵ is a kernel weighting function with bandwidth ϵ.

Alternative methodology

The paper's comparison of BSL and ABC:

- The BSL posterior fell in the vicinity of the ABC posterior for the Ricker example, despite the non-normality of the summary statistics used
- ABC has two tuning parameters (the number of Monte Carlo samples n, and the bandwidth ϵ), where as BSL only has one (n)
- The curse of dimensionality impacts ABC more than BSL due to its nonparametric nature

(We did not implement ABC, since the BSL paper noted it ran ABC for 25 million iterations…)

Reflecting on this paper

We briefly discuss strengths and points of improvements for the BSL paper:

Strengths:

- The authors provided implementations for their algorithm and experiments
  - We found this after implementing the algorithm and examples ourselves
- The examples given appropriately increased in complexity
  - The toy Poisson example used simple statistics and provided an analytical posterior to compare to
  - The Ricker example demonstrated the method's abilities with more complex statistics

Reflecting on this paper

Points of improvement:

- The authors provided implementation in Matlab (kidding…)
- The paper did not thorougly discuss the limitations of BSL, particularly when it came to the choice of statistic
- The paper did not motivate the use of synthetic likelihood in a Bayesian setting very well, beyond *"It is trivial to consider a Bayesian version of this"*

References

- Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*.
- Price, L. F., Drovandi, C. C., Lee, A., & Nott, D. J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*.
- Sisson, S. A. (2011). Likelihood-free markov chain monte carlo. *Handbook of Markov Chain Monte Carlo*.
  - Discusses *approximate Bayesian computation (ABC)*

Thank you

Thank you for listening!