Bayesian Synthetic Likelihood

Thanasi Bakis, Brian Schetzsle

March 16, 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.

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

Additionally, suppose N_t is not actually observed, but a sample from the population $Y_t|N_t \sim \mathit{Poisson}(\phi N_t)$ 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:

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

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.

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:

$$p(\boldsymbol{\theta}|\boldsymbol{s}_{\boldsymbol{Y}})$$

where $\boldsymbol{s}_{\boldsymbol{Y}}$ is a vector of summary statistics for $\boldsymbol{Y}$, eg. the mean, quantiles, etc.

This would use a so-called "synthetic likelihood" $p(\boldsymbol{s}_{\boldsymbol{Y}}|\boldsymbol{\theta})$ 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 otherwords, 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 $\boldsymbol{s}_{\boldsymbol{Y}_i^*}$ in the same manner as $\boldsymbol{s}_{\boldsymbol{Y}}$.

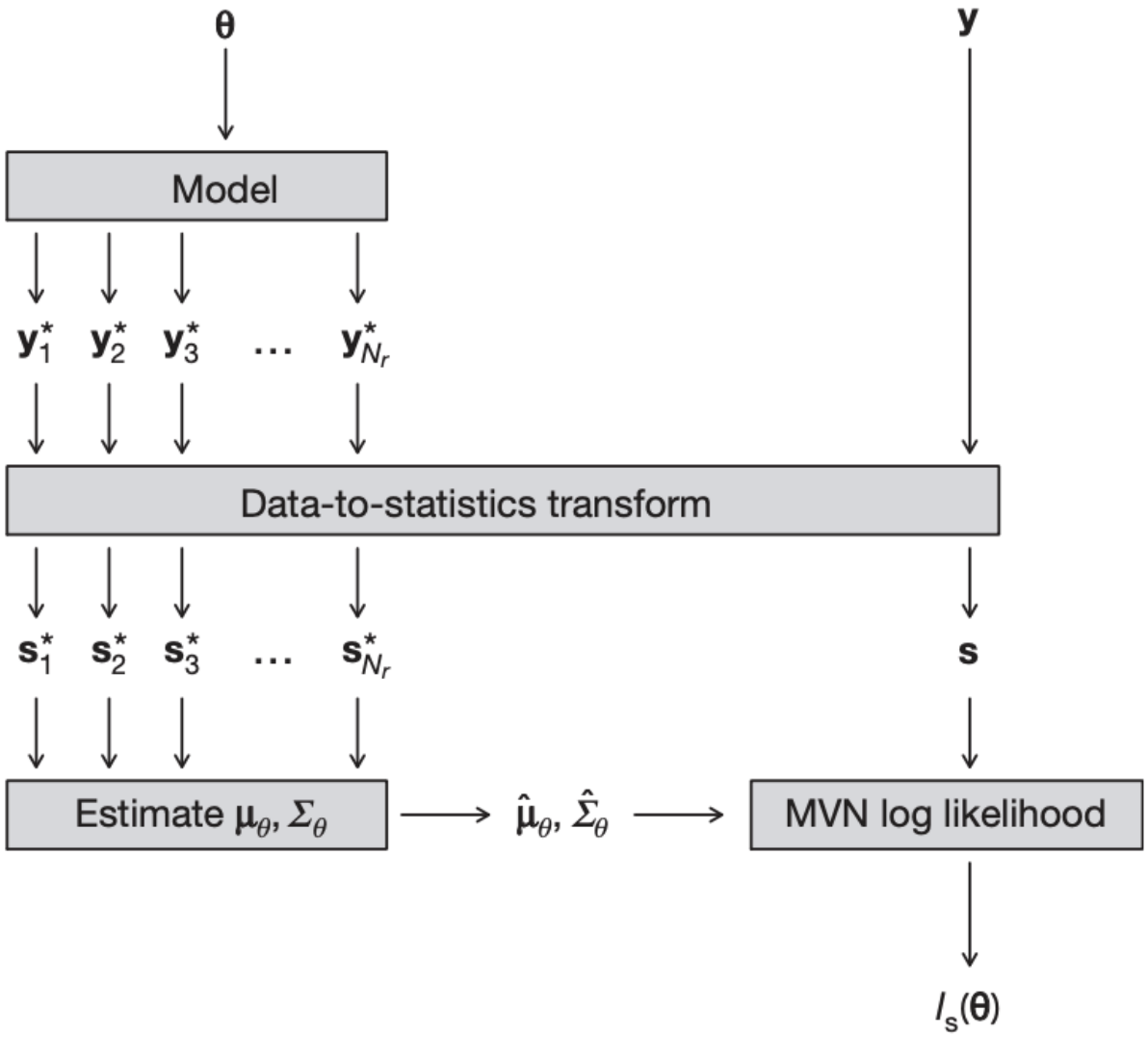
Then,

$$\hat{\boldsymbol{\mu}}_\theta = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{s}_{\boldsymbol{Y}_i^*}$$

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

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

Bayesian synthetic likelihood



A toy example

Consider the model:

$$Y_i|\lambda \overset{\text{iid}}{\sim} Poisson(\lambda) \qquad i = 1, ..., 100$$
$$\lambda \sim Gamma(\alpha = 0.001, \beta = 0.001)$$

Suppose we generate observations $Y_1, ..., Y_{100}$ using $\lambda = 30$ and want to conduct inference on λ.

We want to find $p(\lambda|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\lambda)p(\lambda)$ without evaluating $p(\boldsymbol{Y}|\lambda)$.

A toy example

Note:

$$p(\lambda|\boldsymbol{Y}) \propto \left[\prod_{i=1}^{100} \frac{\lambda^{Y_i}}{Y_i!} e^{-\lambda}\right] \left[\frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}\right]$$

$$\propto \lambda^{\alpha+\sum_{i=1}^{100} Y_i - 1} e^{-\lambda(\beta+n)}$$

$$\sim Gamma(\alpha = 0.001 + \sum_{i=1}^{100} Y_i, \beta = 100.001)$$

So, in this toy example, the posterior distribution is known analytically.

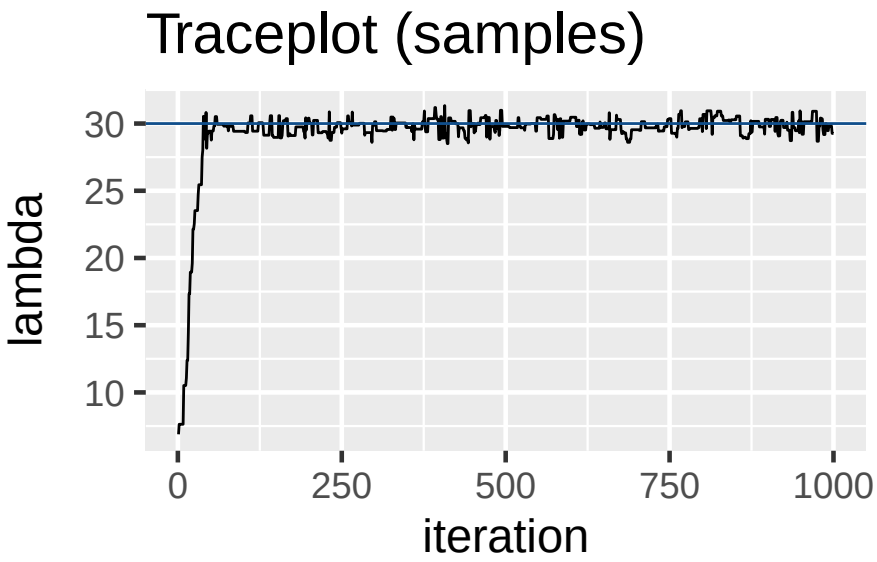
A toy example

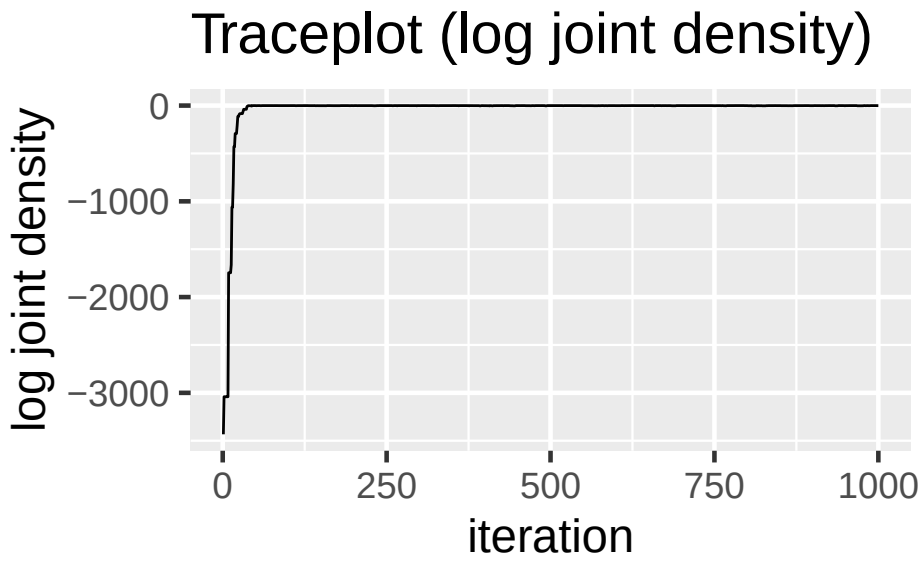
How should we choose a statistic? The paper uses the mean:

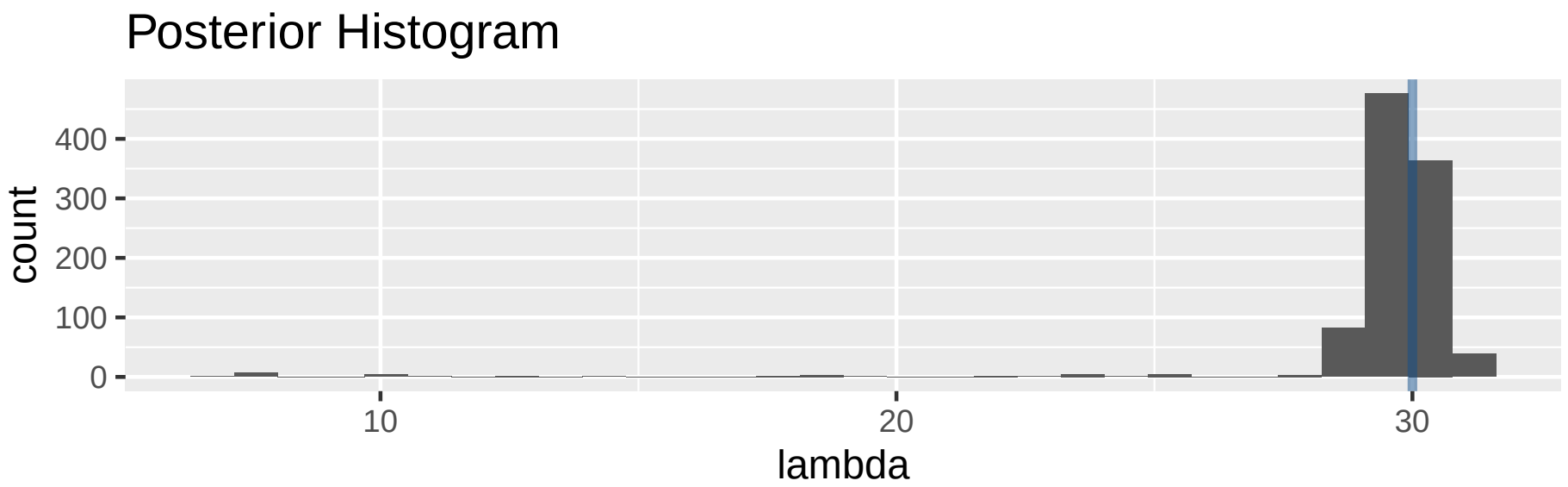
$$s_Y = \frac{1}{100} \sum_{i=1}^{100} Y_i$$

This is the sufficient statistic for the Poisson distribution; all the information contained in the data is also contained in this statistic. Also, by the central limit theorem, the distribution of the mean of a Poisson sample can be adequately approximated by a normal distribution, so synthetic likelihood should perform well in this setting.

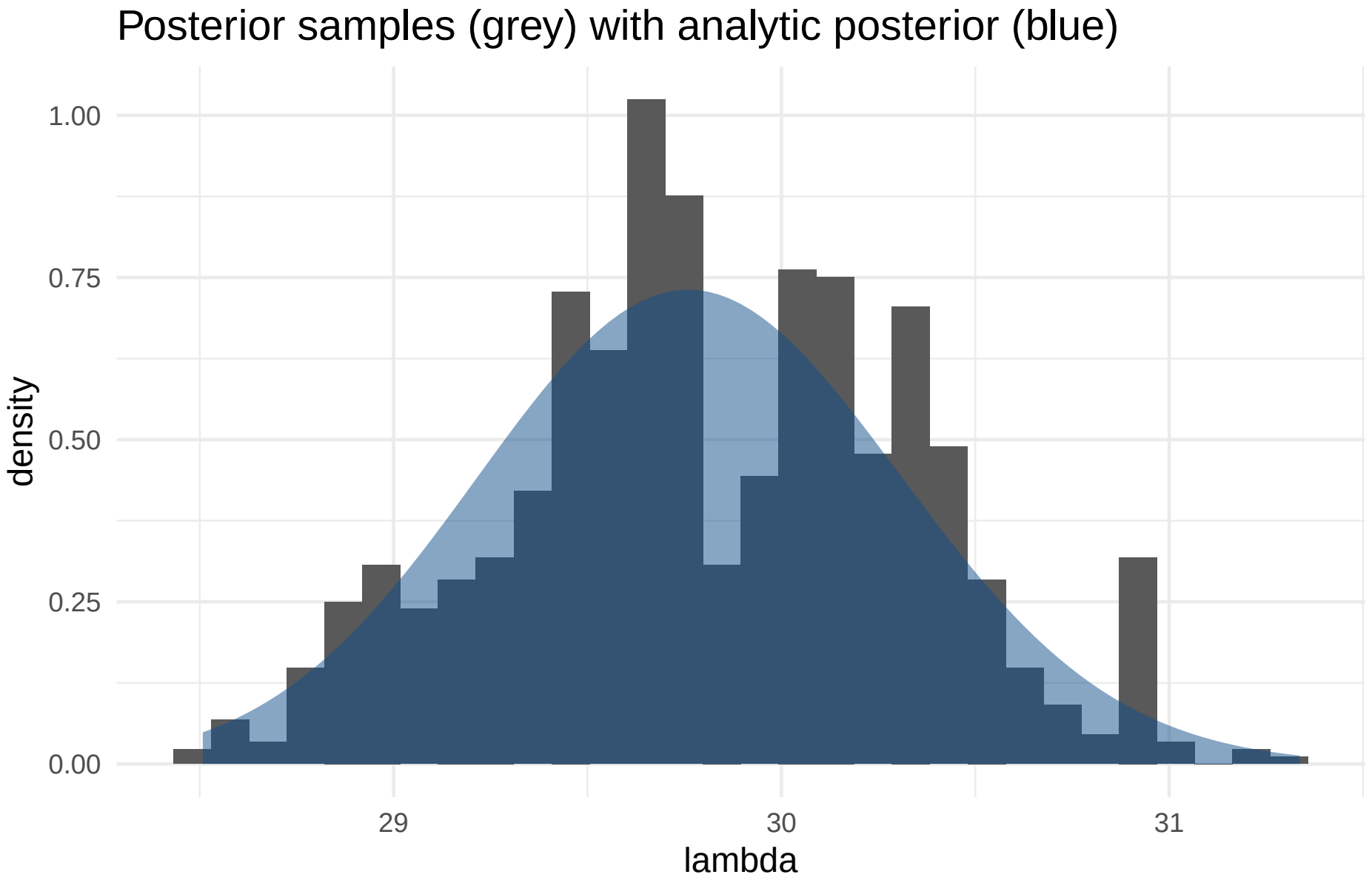
A toy example







A toy example



Further exploration

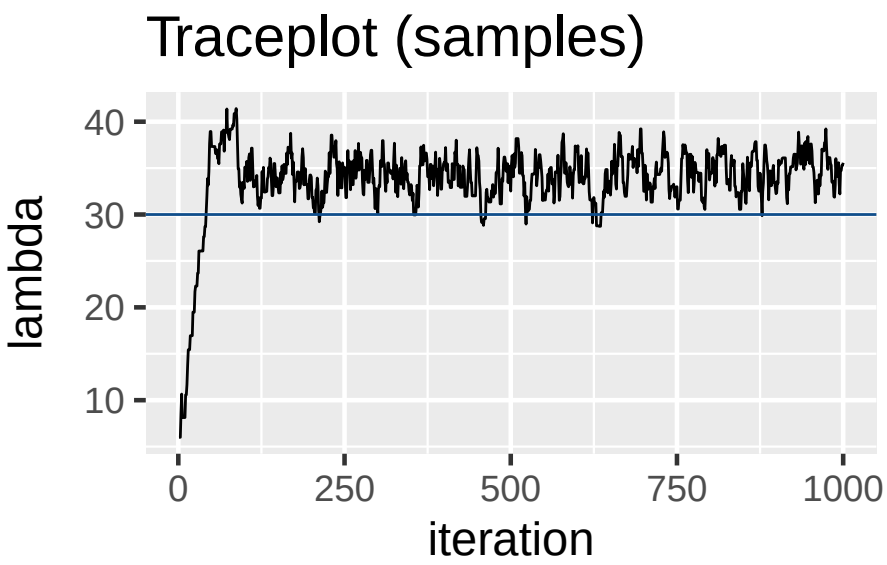
We were interested in trying other statistics to see how well synthetic likelihood performed. This was our own exploration and was not addressed in the paper.

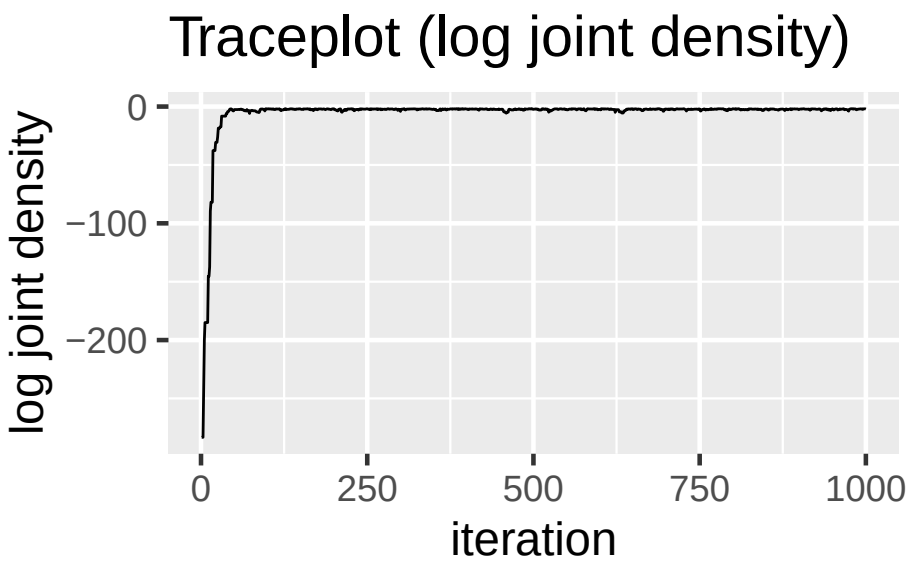
We started with the maximal statistic:

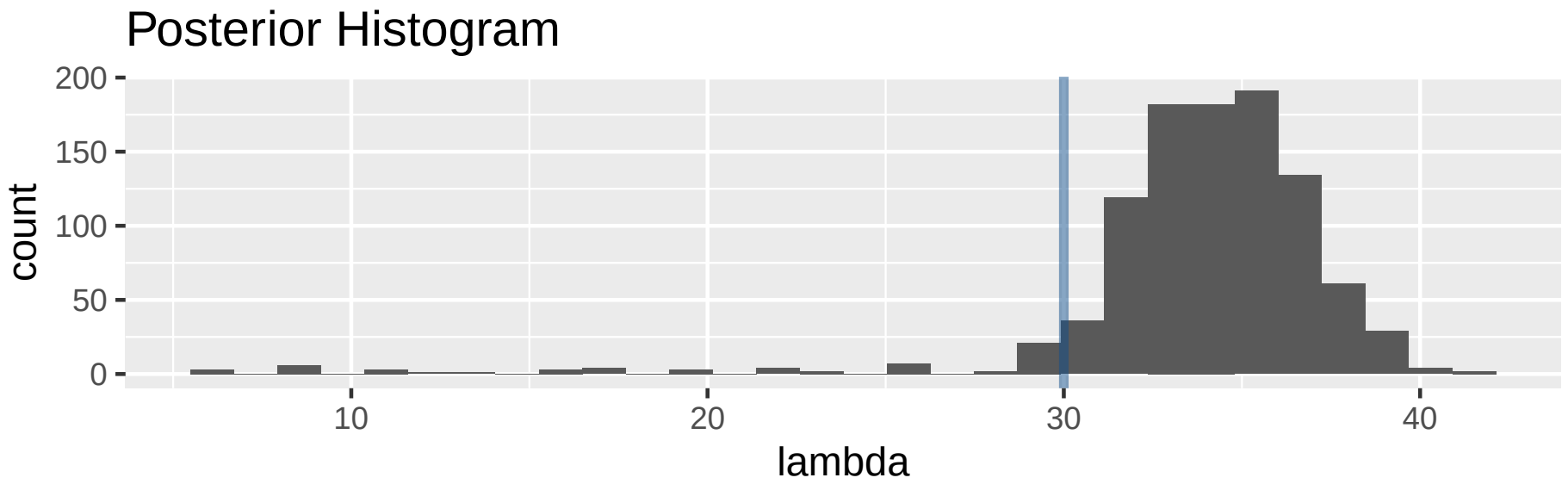
$$s_Y = \max(\boldsymbol{Y})$$

This is not a sufficient statistic for Poisson data and also not approximately normally distributed across many samples. We would expect the synthetic likelihood method to have a harder time identifying the true analytic posterior.

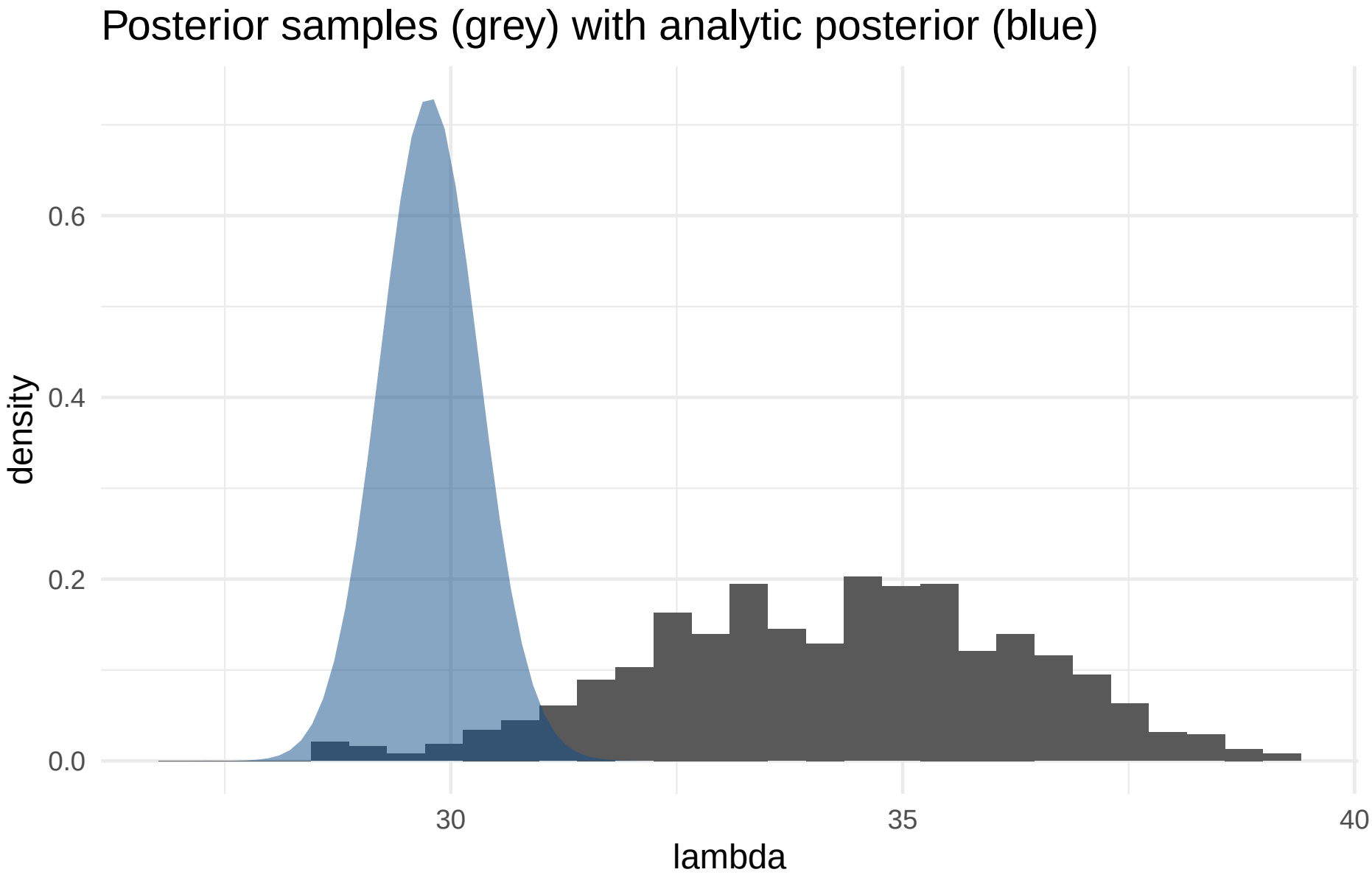
Further Exploration





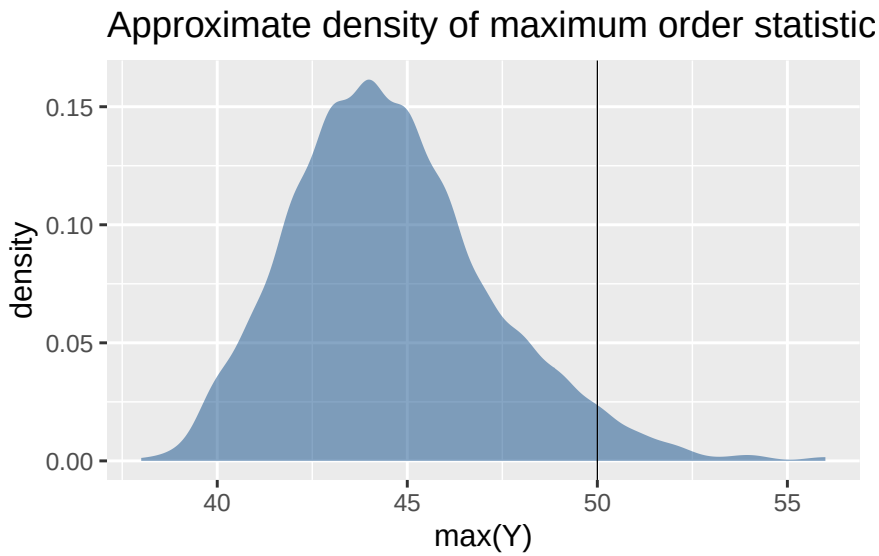


Further exploration



Further exploration

The synthetic likelihood approximation overestimated the true lambda. This is because the data had an unusually large maximal value.

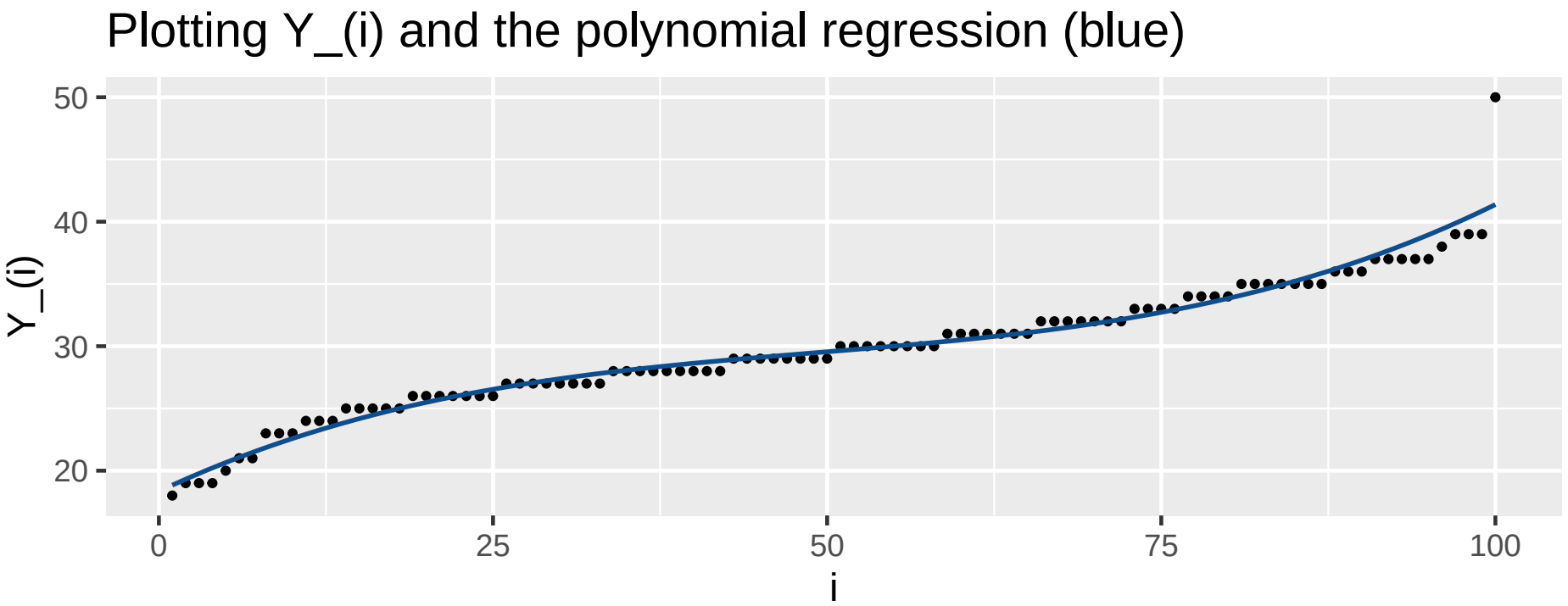


This illustrates how sensitive this method is to the choice of statistic. The approximation will perform optimally on sufficient statistics, but with complicated models, it may not be possible to identify the sufficient statistics. Therefore, statistics must be chosen which capture the important features of the model.

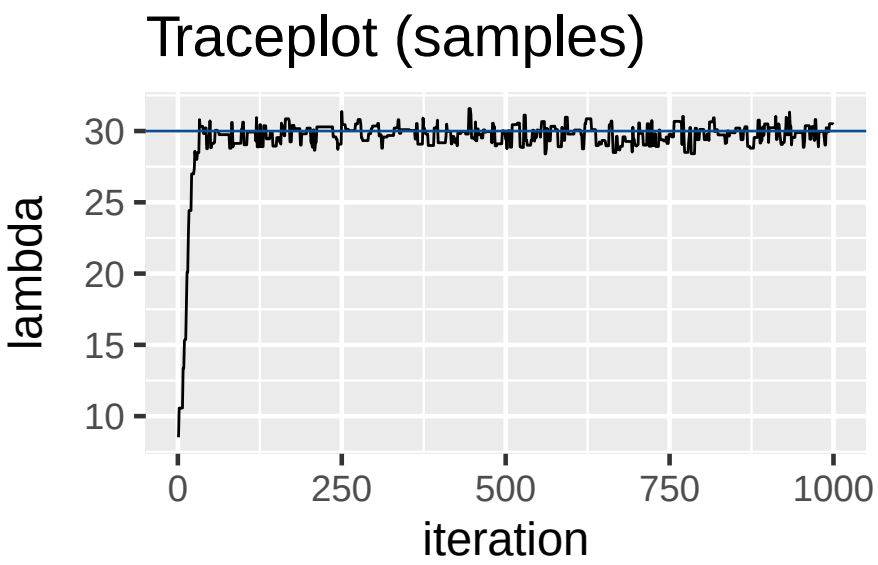
Further exploration

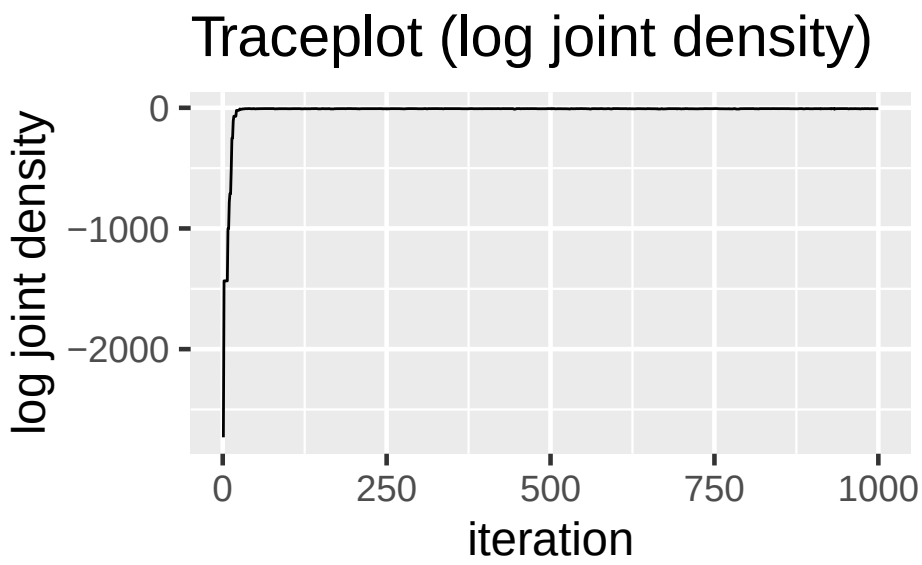
As a final choice of statistics, we used the coefficients of a polynomial regression on the ordered data.

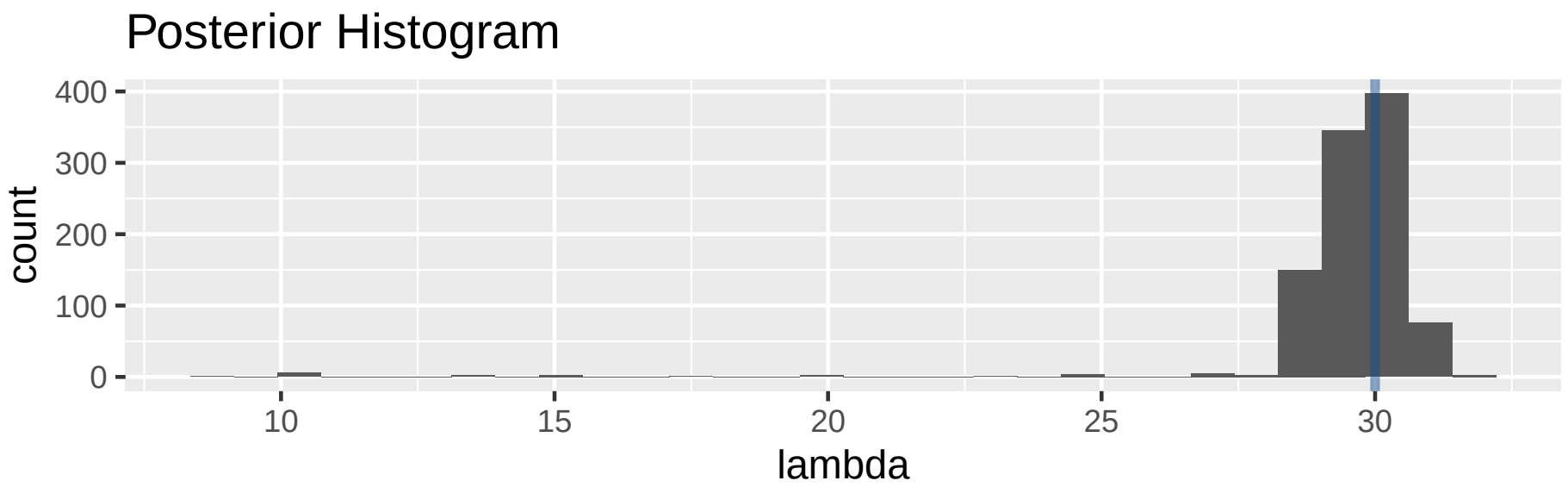
$$Y_{(i)} = \beta_0 + \beta_1 i + \beta_2 i^2 + \beta_3 i^3$$



Further exploration







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:

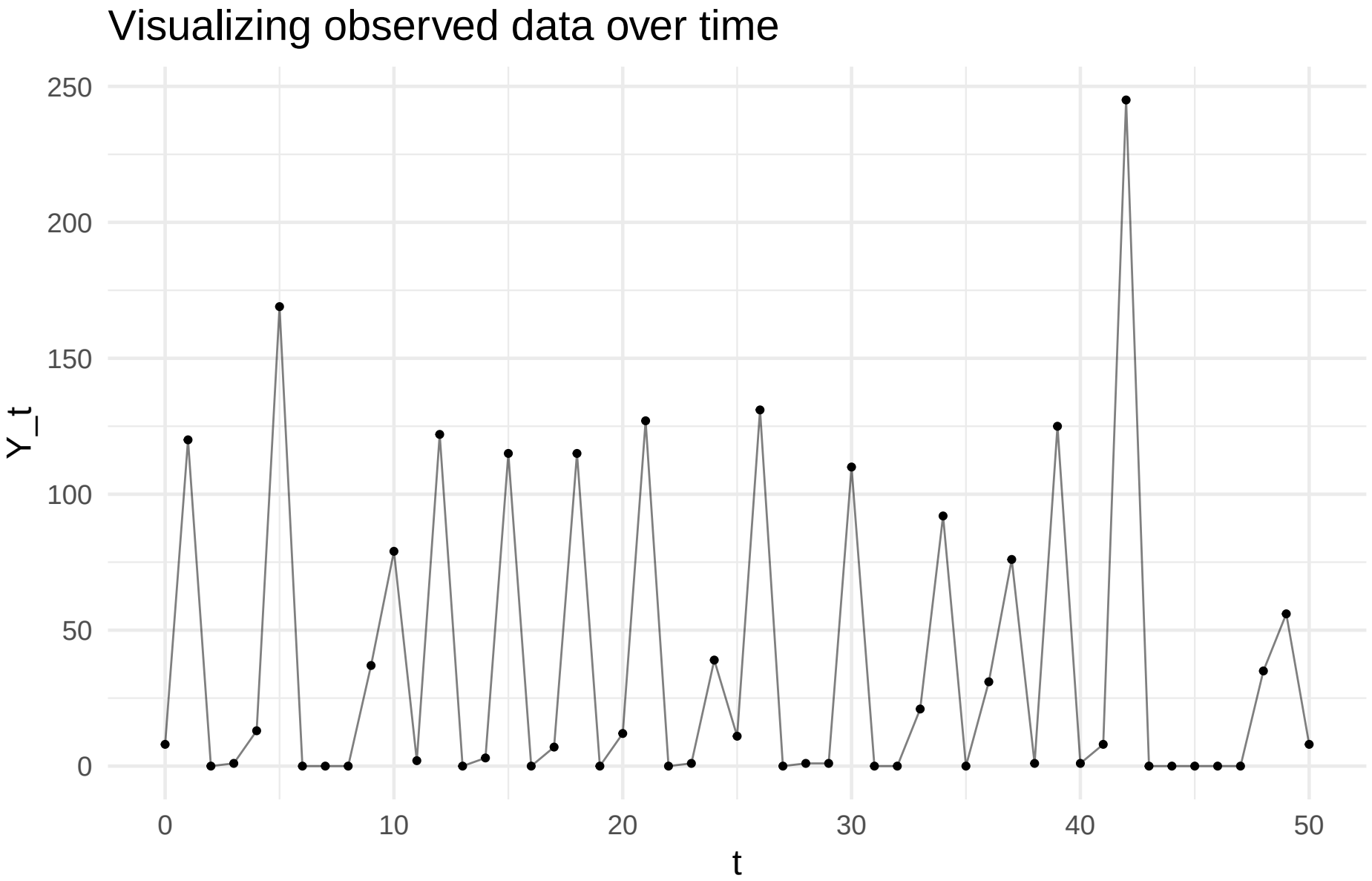
$$N_t|N_{t-1} \sim LogNormal(\log N_{t-1} + \log r - N_{t-1}, \sigma^2)$$
$$Y_t|N_t \sim Poisson(\phi N_t)$$

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

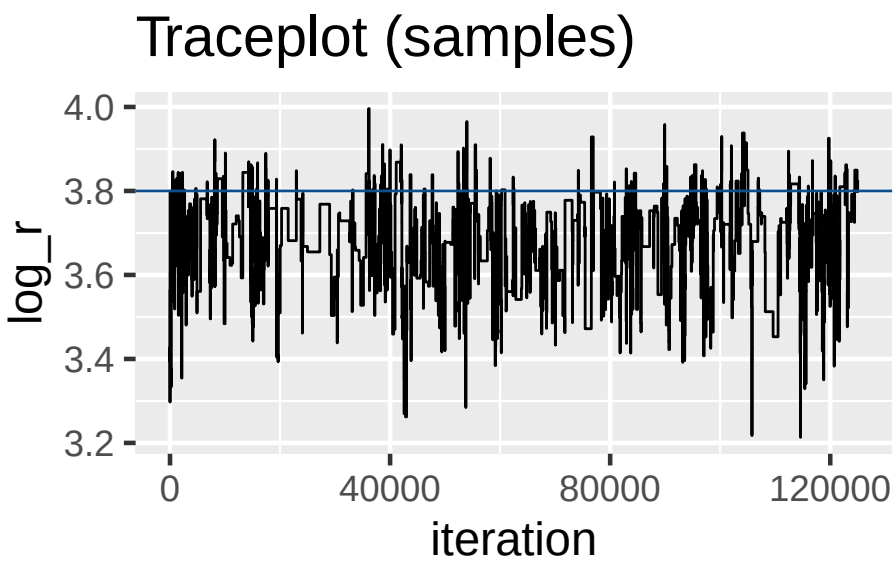


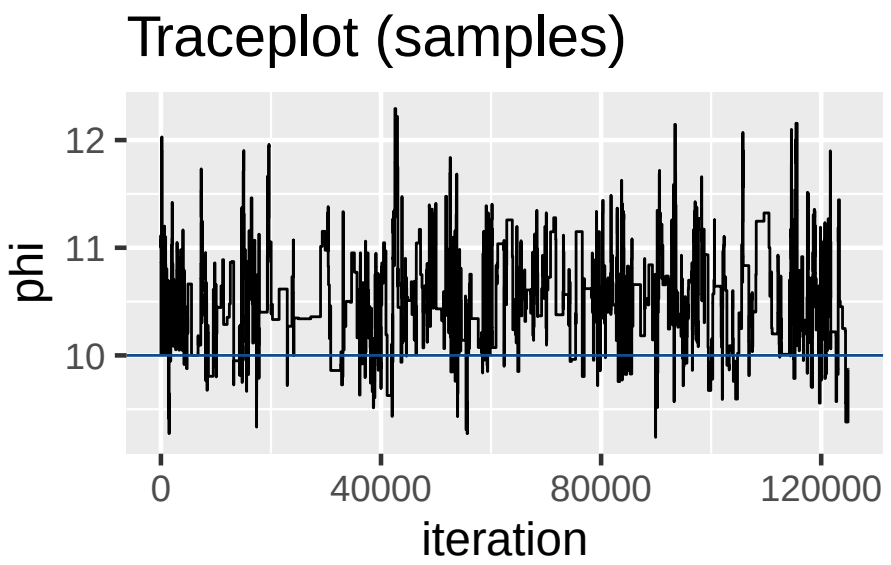
Ricker population model example

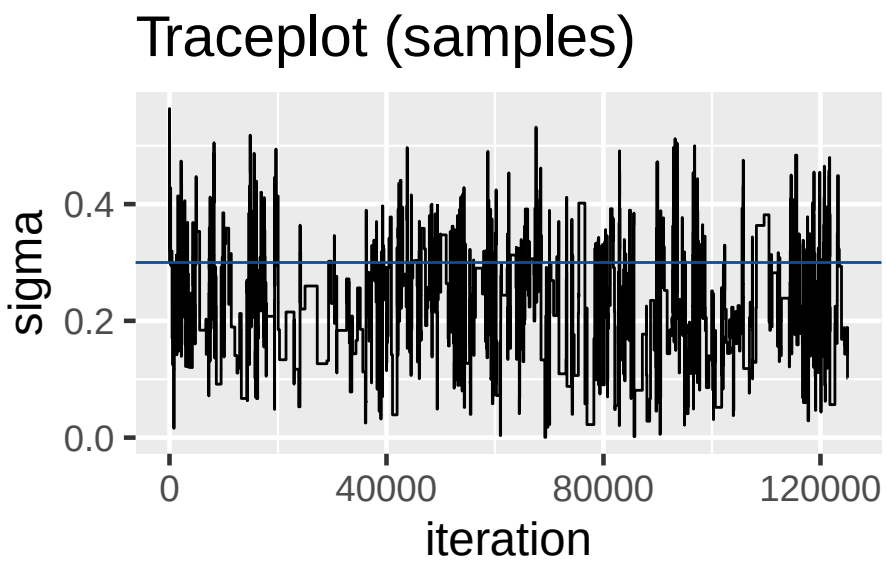
This data is noisy and runs over time. What sorts of statistics will we use?

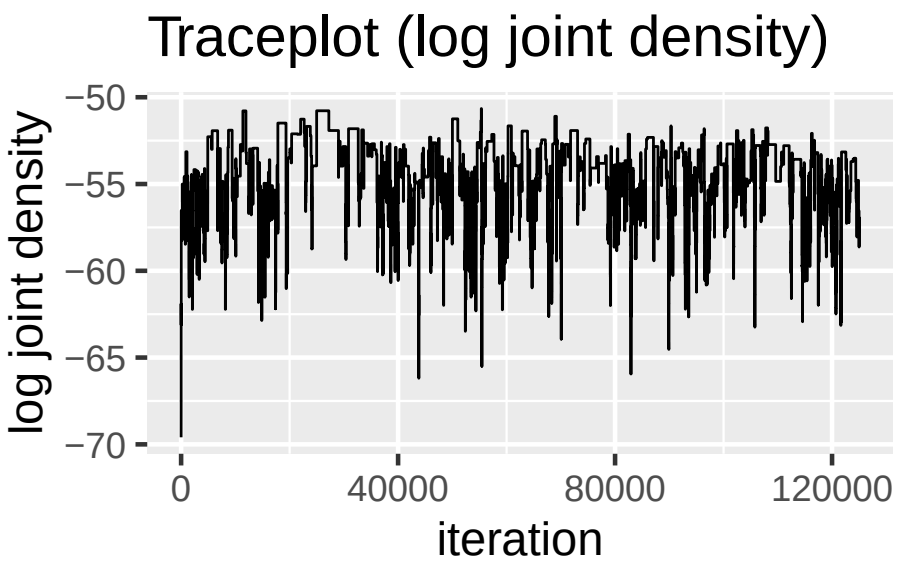
- **Marginal distribution statistics**: to summarize the "shape" of the marginal distribution
  - Mean: $\overline{Y}$
  - Number of zeros: $\sum_{t=1}^{50} \mathbf{1}_{\{0\}}(Y_t)$
- **Dynamic process statistics**: characterize the relationship between Y_t and Y_{t-1} (and possibly more history)
  - 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

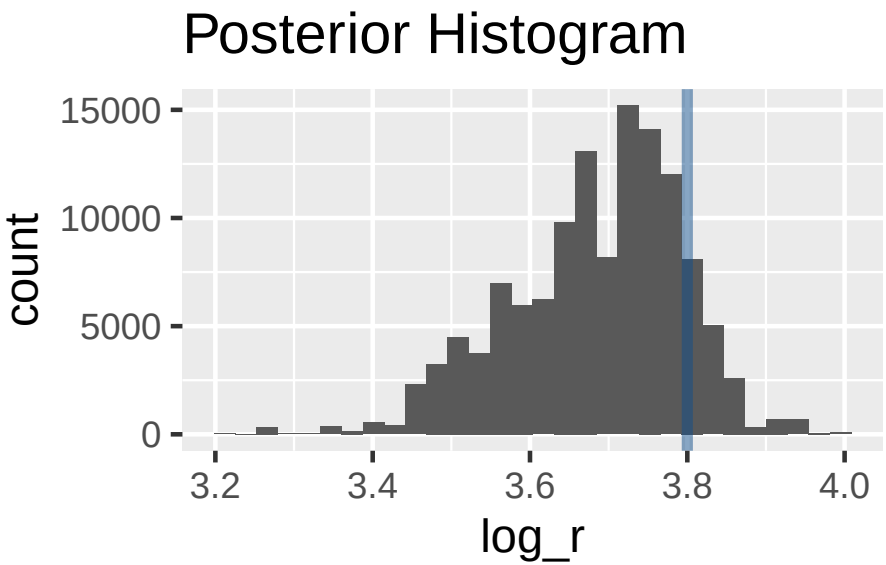


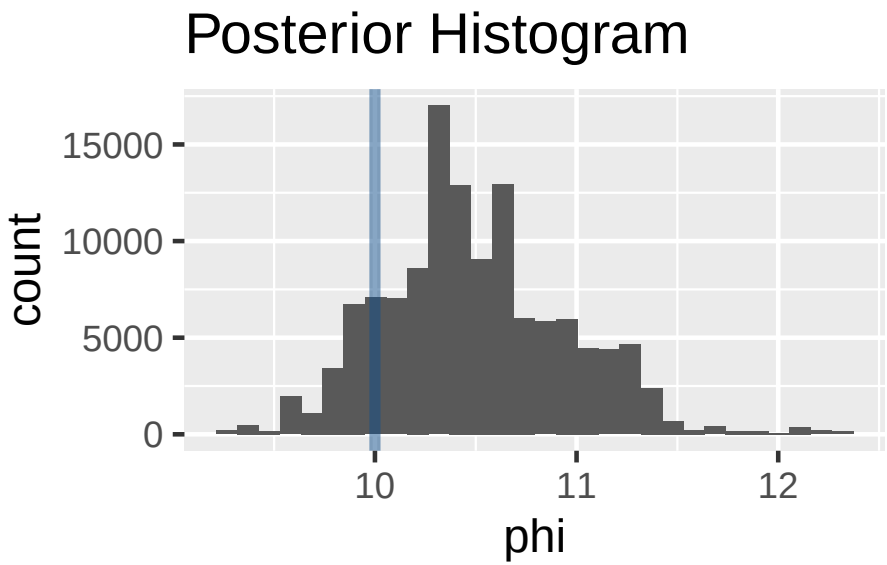


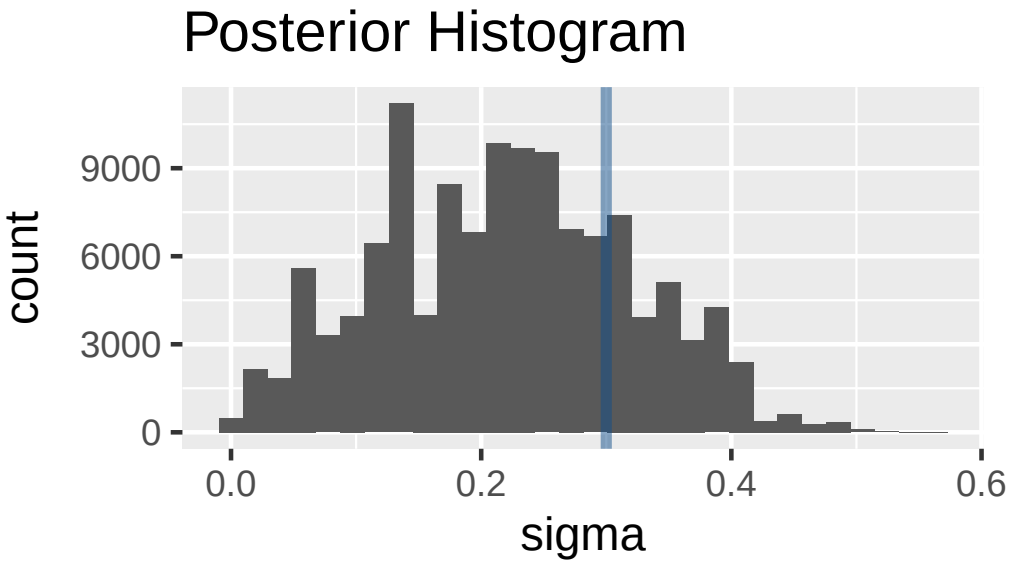




Ricker population model example







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 BSL paper finds that the BSL posterior fell in the vicinity of the ABC posterior for the Ricker example, despite the non-normality of the summary statistics used. It also notes that ABC has two tuning parameters (the number of Monte Carlo samples n, and the bandwidth ϵ), where as BSL only has one (n). Finally, the paper notes that 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!