Choosing Summary Statistics for Approximate Bayesian Compution (ABC)

Nathan Wikle

STAT 540

Project Presentation, 17 April 2018

What is ABC?

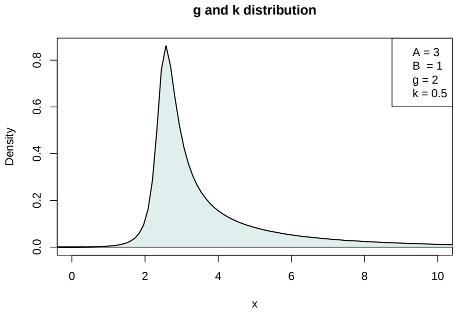
Motivating Problem: How do we perform Bayesian inference when the likelihood function $\ell(\boldsymbol{y}|\boldsymbol{\theta})$ is unavailable?

- e.g., likelihood is given as an intractable integral
  - coalescent models in population genetics
- e.g., intractable normalizing constant
  - Gibbs random fields, point process models, etc

Many Bayesian approaches to inference can no longer be applied!

However, if we can easily simulate from the likelihood,
ABC methods provide an attractive solution.

A Motivating Example



The g and k distribution:

- extension of Normal distribution that accounts for skewness and kurtosis
- CDF and pdf are unavailable in closed form, but the quantile function is given by

$$Q_{gk}(z; A, B, g, k) = A + B\left(1 + 0.8\tanh\left(\frac{gz}{2}\right)\right)z(1+z^2)^k, \quad z \sim N(0,1).$$

Good candidate for ABC: 1) likelihood unavailable, 2) easy to simulate

Rejection-ABC

Some notation:

$\boldsymbol{y}$, the observed data

$\eta(\boldsymbol{y})$, summary statistics of $\boldsymbol{y}$

$\rho > 0$, a distance on η

$\epsilon > 0$, a tolerance level

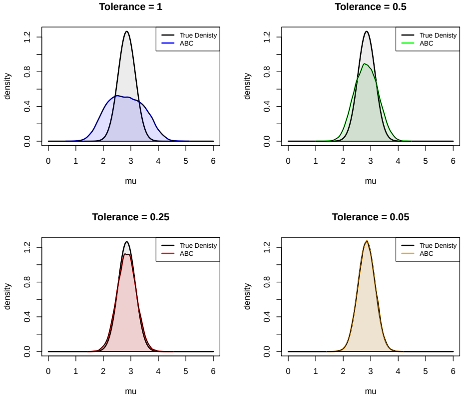
Rejection Algorithm:

- for $i = 1, 2, \ldots, N$:
  - repeat until $\rho\{\eta(\boldsymbol{z}), \eta(\boldsymbol{y})\} \leq \epsilon$
    1. Sample $\boldsymbol{\theta}'$ from $\pi(\cdot)$
    2. Simulate $\boldsymbol{z}$ from $\ell(\cdot|\boldsymbol{\theta}')$
  - set $\boldsymbol{\theta}_i = \boldsymbol{\theta}'$

- The algorithm samples from $\pi_\epsilon(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y})$, a joint posterior distribution of θ and $\boldsymbol{z}$, where $\boldsymbol{z}$ is ϵ-close to $\boldsymbol{y}$.
- **The basic idea of ABC:**

$$\pi_\epsilon(\boldsymbol{\theta}|\boldsymbol{y}) = \int \pi_\epsilon(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{y}) d\boldsymbol{z} \approx \pi(\boldsymbol{\theta}|\boldsymbol{y}).$$

Challenges of ABC



- The success of ABC is dependent on the choice of calibration parameters.
- Optimal if η is sufficient and $\epsilon \to 0$.
- In practice, η is not sufficient, and small $\epsilon =$ larger computational time.

In general, the ABC literature is focused on:

1) choice of appropriate calibration parameters

2) efficient sampling algorithms

e.g., ABC-MCMC, ABC-SMC, etc.

Choosing Appropriate Summary Statistics

Choosing η is challenging

- problem specific
- sufficient statistics are the gold standard
- want $\dim(\eta)$ as close to $\dim(\boldsymbol{\theta})$ as possible

Three common classes of methods (Blum et al., 2013):

1) best subset selection (Joyce and Marjoram, 2008)
2) post-processing (Beaumont et al., 2002; Blum and Francois, 2010)
3) semi-automatic ABC (Fearnhead and Prangle, 2012)

Semi-Automatic ABC (Fearnhead and Prangle, 2012)

Idea: Assume interest is in point estimates of model parameters

- If $\boldsymbol{\theta}_0$ is the true parameter value, and $\hat{\boldsymbol{\theta}}$ is an estimate, choose η that minimizes

$$L(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}}) = (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})' A (\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})$$

- $L(\boldsymbol{\theta}_0, \hat{\boldsymbol{\theta}})$ is minimized if $\eta(\boldsymbol{y}) = E(\boldsymbol{\theta}|\boldsymbol{y})$
- Resulting $\eta(\cdot)$ is low-dimensional
- Turned our problem into finding $\eta(\boldsymbol{y}) \approx E(\boldsymbol{\theta}|\boldsymbol{y})$
- Advantage: can be applied to any ABC algorithm

Semi-automatic ABC:

1) Simulate many $(\boldsymbol{\theta}, \boldsymbol{z})$ "pilot" values

   - Simulate $\boldsymbol{\theta} \sim \pi(\cdot)$ and $\boldsymbol{z} \sim \ell(\cdot, \boldsymbol{\theta})$

2) Estimate $\eta(\boldsymbol{z}) \approx E(\boldsymbol{\theta}|\boldsymbol{z})$

3) Use $\eta(\boldsymbol{z})$ as summary statistic for ABC

More on Semi-Automatic ABC

The authors use **linear regression** on the simulated $\{(\boldsymbol{\theta}, \boldsymbol{z})\}$ to estimate $E(\boldsymbol{\theta}|\boldsymbol{z})$.

- $\theta_i = E(\theta_i|\boldsymbol{z}) + \epsilon_i = \beta_0^{(i)} + \boldsymbol{\beta}^{(i)} f(\boldsymbol{z}) + \epsilon_i$, for each θ_i.

My idea: What if we use regularization methods and nonlinear models to estimate $E(\boldsymbol{\theta}|\boldsymbol{z})$? In particular, I considered:

- **LASSO:** minimize $RSS + \lambda \sum_{j=1}^{p} |\beta_j|$
- **Ridge:** minimize $RSS + \lambda \sum_{j=1}^{p} \beta_j^2$
- **Random Forests:** bootstrap aggregation of regression trees

Note: Although the authors don't discuss these extensions, regularization and nonlinear models have been used in post-processing of ABC data for some time (e.g., Beaumont et al. (2002), Blum and Francois (2010), Blum et al. (2013)).

My Work

Motivation:

- Methods are easy to implement in R.
- May lead to better predictions than OLS solution
- May avoid overfitting the initial pilot run.
- In some cases, can help deal with collinearity in $f(\boldsymbol{z})$.
- Can handle large number of covariates.

I compared the performance using two examples:

- A toy example: Normal likelihood with conjugate priors
  - possible to compare performance to the true posterior
  - less computational cost, used Rejection-ABC
- The g and k distribution:
  - true posterior is not available; compare to results in paper
  - computing cost is noticeable, used ABC-MCMC

Normal Toy Example

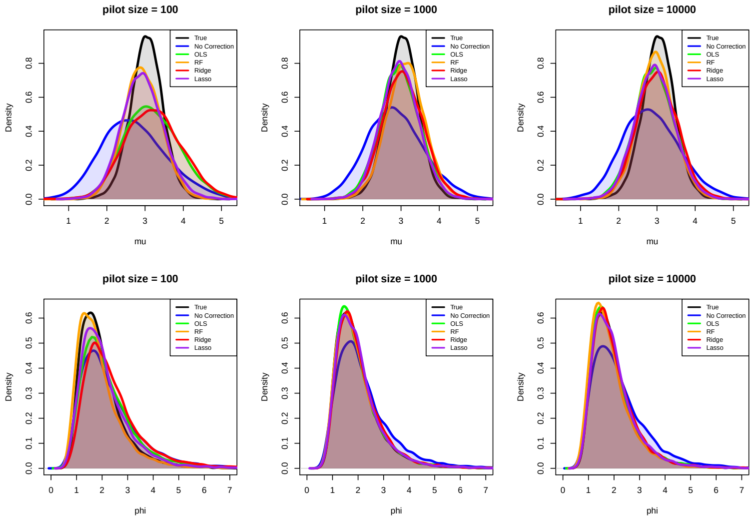
$$\boldsymbol{y} \overset{iid}{\sim} N(\mu, \sigma^2), \quad \mu|\sigma^2 \sim N(\mu_0, \frac{\sigma^2}{p_0}), \quad \frac{1}{\sigma^2} \sim \Gamma(\nu_0/2, s_0/2)$$

- Note: this is a conjugate prior, and we can sample directly from the posterior for comparison.
- $f(\boldsymbol{z})$ contains the true sufficient statistics $[\bar{\boldsymbol{z}}, SS(\boldsymbol{z})]$ and the median.
  - Also included: transformations of these values, and 15 $N(0,1)$ covariates.

Questions to consider:

- Approximation to the true posterior?
- Computational costs?
- How many pilot samples are needed?

Results: Normal Toy Example



g and k distribution

$$\boldsymbol{y} \overset{iid}{\sim} Q_{gk}(\cdot, A, B, g, k), \quad (A, B, g, k) \sim (0, 10)^4$$

- $\boldsymbol{y}_{obs}$ consists of 10,000 observations from $Q_{gk}(\cdot, 3, 1, 2, 0.5)$
- $f(\boldsymbol{z})$ is a vector of 60 equally spaced order statistics and their powers (up to the 4th power)
- ABC - MCMC was used to facilitate sampling

Some notes about implementation:

- each pilot run consisted of 10,000 samples
- semi-automatic ABC-MCMC was implemented for OLS, lasso, ridge, and random forest, 25 times each
- MSE (with respect to the posterior mean) was used to compare methods

Results: g and k distribution

Posterior Mean MSE:

	A	B	g	k
OLS Reg.	**0.000103**	**0.000797**	**0.062392**	0.135695
Lasso	0.001441	0.004164	0.248754	0.558078
Ridge Reg.	0.001451	0.004888	1.156954	0.871882
Random Forests*	0.001405	0.021403	11.666854	**0.008647**

Some observations:

- The penalty term for both lasso and ridge regression was chosen (using CV) to be very small.
- Random forests takes longer to predict than the regression models - significantly increased computational burden of ABC-MCMC.
- ABC-MCMC required a lot of tuning to run well. Perhaps a different ABC algorithm would be preferred.

Conclusions and Future Work

Conclusions:

- Semi-automatic ABC provides a general approach for finding summary statistics for ABC.
- Lasso, ridge regression, and random forests are competitive with OLS regression for choosing η, and should be considered as alternative methods, especially when simulation is costly.

Future Work:

- I hope to compare the success of these approaches on a class of repulsive point processes, presented by Shirota and Gelfand (2017).
- Additional questions to investigate:

  1) How well does semi-automatic ABC perform in ABC-SMC?
  2) Would other nonlinear models (e.g., neural nets) outperform regression and random forests?