Approximate Bayesian Computation (ABC)

David Nott

Department of Statistics and Applied Probability
National University of Singapore

SIAM Conference on Uncertainty Quantification, 2018

Bayesian inference and the basic ABC idea

Data y to be observed, unknowns θ. Construct a model for (y, θ):

$$p(y,\theta) = \overbrace{p(\theta)}^{\text{prior}}\ \overbrace{p(y|\theta)}^{\text{data model}}$$

Condition on y_{obs}, the observed y:

$$\underbrace{p(\theta|y_{\text{obs}})}_{\text{posterior}} \propto p(\theta)p(y_{\text{obs}}|\theta)$$

Summarizing the posterior (calculating probabilities, moments, etc.) - usually done by Markov chain Monte Carlo.

Being able to calculate the likelihood $p(y_{\text{obs}}|\theta)$ seems like a basic requirement ...

Bayesian inference and the basic ABC idea

Data y to be observed, unknowns θ. Construct a model for (y, θ):

$$p(y,\theta) = \overbrace{p(\theta)}^{\text{prior}} \; \overbrace{p(y|\theta)}^{\text{data model}}$$

Condition on y_{obs}, the observed y:

$$\underbrace{p(\theta|y_{\text{obs}})}_{\text{posterior}} \propto p(\theta)p(y_{\text{obs}}|\theta)$$

Summarizing the posterior (calculating probabilities, moments, etc.) - usually done by Markov chain Monte Carlo.

Being able to calculate the likelihood $p(y_{\text{obs}}|\theta)$ seems like a basic requirement ...

Bayesian inference and the basic ABC idea

Data y to be observed, unknowns θ. Construct a model for (y, θ):

$$p(y,\theta) = \overbrace{p(\theta)}^{\text{prior}} \; \overbrace{p(y|\theta)}^{\text{data model}}$$

Condition on y_{obs}, the observed y:

$$\underbrace{p(\theta|y_{\text{obs}})}_{\text{posterior}} \propto p(\theta)p(y_{\text{obs}}|\theta)$$

Summarizing the posterior (calculating probabilities, moments, etc.) - usually done by Markov chain Monte Carlo.

Being able to calculate the likelihood $p(y_{\text{obs}}|\theta)$ seems like a basic requirement ...

Bayesian inference and the basic ABC idea

Data y to be observed, unknowns θ. Construct a model for (y, θ):

$$p(y,\theta) = \overbrace{p(\theta)}^{\text{prior}} \; \overbrace{p(y|\theta)}^{\text{data model}}$$

Condition on y_{obs}, the observed y:

$$\underbrace{p(\theta|y_{\text{obs}})}_{\text{posterior}} \propto p(\theta)p(y_{\text{obs}}|\theta)$$

Summarizing the posterior (calculating probabilities, moments, etc.) - usually done by Markov chain Monte Carlo.

Being able to calculate the likelihood $p(y_{\text{obs}}|\theta)$ seems like a basic requirement ...

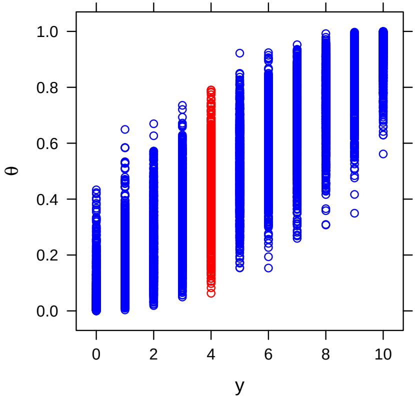
The basic ABC idea

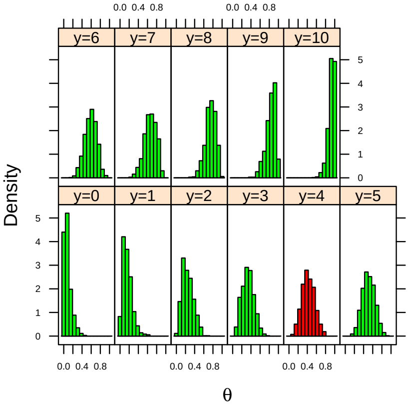
Rubin (1984) explaining Bayes rule

"*How is the conceptual content of this theorem easily conveyed? ... Suppose we first draw equally likely values of θ from $p(\theta)$ ... For each θ_j we now draw an X from $f(X|\theta = \theta_j)$; ... Suppose we collect together all X_j that match the observed X and then all θ_j that correspond to these X_j ... formally, this collection of θ values represents the posterior distribution of θ.*"

The basic ABC idea

Toy example: $y|\theta$ Binomial$(10, \theta)$, $\theta \sim$ Beta$(1, 1)$, $y_{\text{obs}} = 4$





Replacing exact conditioning with "good enough"

- Let $d(\cdot,\cdot)$ be a distance defined in the data space and $\epsilon > 0$ be a tolerance.
- The basic rejection ABC algorithm is (Pritchard *et al.*, 1999):

Generate a joint sample (θ, y) from the model until $d(y, y_{\text{obs}}) < \epsilon$, keeping the θ sample when this occurs.

- The output θ is an approximate draw from $p(\theta|y)$.
- The distance is usually constructed by reducing y to a summary $S = S(y) \sim p(S|\theta)$ informative about θ and then $d(\cdot,\cdot)$ is defined in the space of summary statistics.
- If S is sufficient and $\epsilon \to 0$ the above algorithm is exact.

Replacing exact conditioning with "good enough"

- Let $d(\cdot,\cdot)$ be a distance defined in the data space and $\epsilon > 0$ be a tolerance.
- The basic rejection ABC algorithm is (Pritchard *et al.*, 1999):

Generate a joint sample (θ, y) from the model until $d(y, y_{\text{obs}}) < \epsilon$, keeping the θ sample when this occurs.

- The output θ is an approximate draw from $p(\theta|y)$.
- The distance is usually constructed by reducing y to a summary $S = S(y) \sim p(S|\theta)$ informative about θ and then $d(\cdot,\cdot)$ is defined in the space of summary statistics.
- If S is sufficient and $\epsilon \to 0$ the above algorithm is exact.

Replacing exact conditioning with "good enough"

- Let $d(\cdot,\cdot)$ be a distance defined in the data space and $\epsilon > 0$ be a tolerance.
- The basic rejection ABC algorithm is (Pritchard *et al.*, 1999):

> Generate a joint sample (θ, y) from the model until $d(y, y_{\text{obs}}) < \epsilon$, keeping the θ sample when this occurs.

- The output θ is an approximate draw from $p(\theta|y)$.
- The distance is usually constructed by reducing y to a summary $S = S(y) \sim p(S|\theta)$ informative about θ and then $d(\cdot,\cdot)$ is defined in the space of summary statistics.
- If S is sufficient and $\epsilon \to 0$ the above algorithm is exact.

Replacing exact conditioning with "good enough"

- Let $d(\cdot,\cdot)$ be a distance defined in the data space and $\epsilon > 0$ be a tolerance.
- The basic rejection ABC algorithm is (Pritchard *et al.*, 1999):

Generate a joint sample (θ, y) from the model until $d(y, y_{\text{obs}}) < \epsilon$, keeping the θ sample when this occurs.

- The output θ is an approximate draw from $p(\theta|y)$.
- The distance is usually constructed by reducing y to a summary $S = S(y) \sim p(S|\theta)$ informative about θ and then $d(\cdot,\cdot)$ is defined in the space of summary statistics.
- If S is sufficient and $\epsilon \to 0$ the above algorithm is exact.

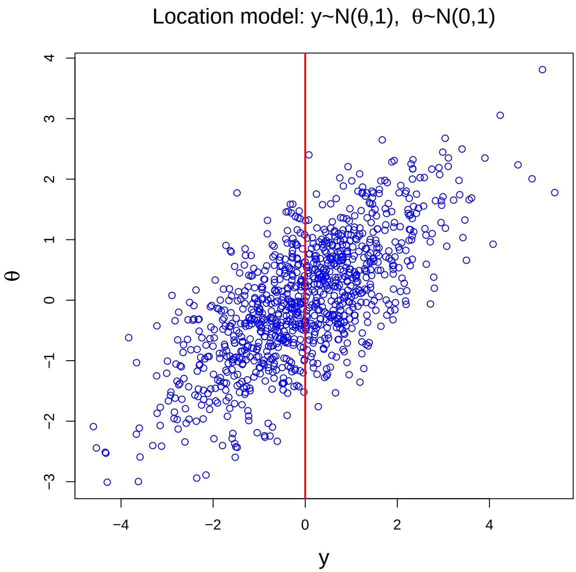
Replacing exact conditioning with "good enough"

- Let $d(\cdot,\cdot)$ be a distance defined in the data space and $\epsilon > 0$ be a tolerance.
- The basic rejection ABC algorithm is (Pritchard *et al.*, 1999):

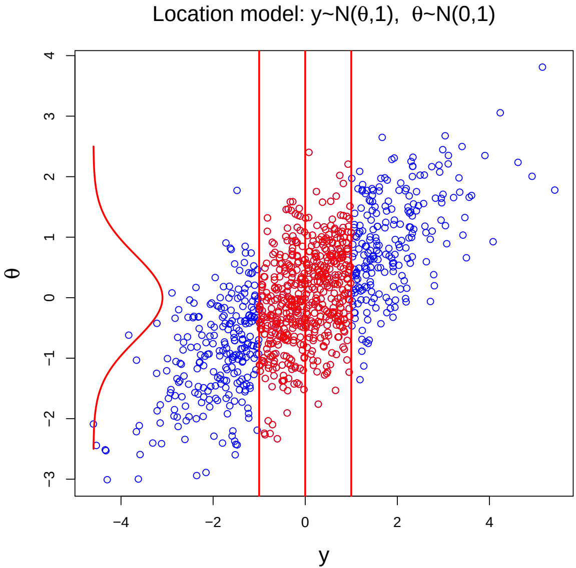
Generate a joint sample (θ, y) from the model until $d(y, y_{\text{obs}}) < \epsilon$, keeping the θ sample when this occurs.

- The output θ is an approximate draw from $p(\theta|y)$.
- The distance is usually constructed by reducing y to a summary $S = S(y) \sim p(S|\theta)$ informative about θ and then $d(\cdot,\cdot)$ is defined in the space of summary statistics.
- If S is sufficient and $\epsilon \to 0$ the above algorithm is exact.

Bayesian inference and the basic ABC idea



Bayesian inference and the basic ABC idea



Why ABC?

- The basic rejection ABC algorithm only required simulations $(\theta, y) \sim p(\theta)p(y|\theta)$.
- We don't need to compute the likelihood $p(y_{\text{obs}}|\theta)$.
- Difficulties:
  - Choosing the distance measure (including the summaries S)
  - Choice of tolerance ϵ
  - Validating the approximation

Why ABC?

- The basic rejection ABC algorithm only required simulations $(\theta, y) \sim p(\theta)p(y|\theta)$.
- We don't need to compute the likelihood $p(y_{\text{obs}}|\theta)$.
- Difficulties:
  - Choosing the distance measure (including the summaries S)
  - Choice of tolerance ϵ
  - Validating the approximation

Why ABC?

- The basic rejection ABC algorithm only required simulations $(\theta, y) \sim p(\theta)p(y|\theta)$.
- We don't need to compute the likelihood $p(y_{obs}|\theta)$.
- Difficulties:
  - Choosing the distance measure (including the summaries S)
  - Choice of tolerance ϵ
  - Validating the approximation

Why ABC?

- The basic rejection ABC algorithm only required simulations $(\theta, y) \sim p(\theta)p(y|\theta)$.
- We don't need to compute the likelihood $p(y_{\text{obs}}|\theta)$.
- Difficulties:
  - Choosing the distance measure (including the summaries S)
  - Choice of tolerance ϵ
  - Validating the approximation

Why ABC?

- The basic rejection ABC algorithm only required simulations $(\theta, y) \sim p(\theta)p(y|\theta)$.
- We don't need to compute the likelihood $p(y_{\text{obs}}|\theta)$.
- Difficulties:
  - Choosing the distance measure (including the summaries S)
  - Choice of tolerance ϵ
  - Validating the approximation

Why ABC?

- The basic rejection ABC algorithm only required simulations $(\theta, y) \sim p(\theta)p(y|\theta)$.
- We don't need to compute the likelihood $p(y_{\text{obs}}|\theta)$.
- Difficulties:
  - Choosing the distance measure (including the summaries S)
  - Choice of tolerance ϵ
  - Validating the approximation

Ricker model

Ricker, 1954, Wood, 2010

- Let $S^{(t)}$ be the size of an animal population at time t, $t = 1, \dots, T$.

Ricker model

$$\log S^{(t)} = \log r + \log S^{(t-1)} - S^{(t-1)} + \sigma e^{(t)},$$

$$y^{(t)} | S^{(t)} \sim \text{Poisson}(\psi S^{(t)}).$$

- Parameters $\theta = (\log r, \sigma, \psi)$. Here:
  - r is a growth rate,
  - σ is the standard deviation of environmental noise,
  - ψ is a scale parameter
- Priors: $\log r \sim U[2, 5]$, $\log \sigma \sim U[-2.3, -3]$, $\log \psi \sim U[-1.79, 1.61]$.

Ricker model

Ricker, 1954, Wood, 2010

- Let $S^{(t)}$ be the size of an animal population at time t, $t = 1, \ldots, T$.

Ricker model

$$\log S^{(t)} = \log r + \log S^{(t-1)} - S^{(t-1)} + \sigma e^{(t)},$$

$$y^{(t)} | S^{(t)} \sim \text{Poisson}(\psi S^{(t)}).$$

- Parameters $\theta = (\log r, \sigma, \psi)$. Here:
  - r is a growth rate,
  - σ is the standard deviation of environmental noise,
  - ψ is a scale parameter
- Priors: $\log r \sim U[2, 5]$, $\log \sigma \sim U[-2.3, -3]$, $\log \psi \sim U[-1.79, 1.61]$.

Ricker model

Ricker, 1954, Wood, 2010

- Let $S^{(t)}$ be the size of an animal population at time t, $t = 1, \dots, T$.

Ricker model

$$\log S^{(t)} = \log r + \log S^{(t-1)} - S^{(t-1)} + \sigma e^{(t)},$$

$$y^{(t)} | S^{(t)} \sim \text{Poisson}(\psi S^{(t)}).$$

- Parameters $\theta = (\log r, \sigma, \psi)$. Here:
  - r is a growth rate,
  - σ is the standard deviation of environmental noise,
  - ψ is a scale parameter
- Priors: $\log r \sim U[2, 5]$, $\log \sigma \sim U[-2.3, -3]$, $\log \psi \sim U[-1.79, 1.61]$.

Ricker model

Ricker, 1954, Wood, 2010

- Wood (2010) considered this stochastic version of the Ricker model in developing his synthetic likelihood.
- Motivation: hard to explore some parts of the parameter space in near chaotic models with low noise in the state model with full likelihood methods.
- Consider inference based on summary statistics for which the summary statistic likelihood is better behaved (Fasiolo, Pya and Wood, 2016).

Ricker model

Ricker, 1954, Wood, 2010

- Wood (2010) considered this stochastic version of the Ricker model in developing his synthetic likelihood.
- Motivation: hard to explore some parts of the parameter space in near chaotic models with low noise in the state model with full likelihood methods.
- Consider inference based on summary statistics for which the summary statistic likelihood is better behaved (Fasiolo, Pya and Wood, 2016).

Ricker model

Ricker, 1954, Wood, 2010

- Wood (2010) considered this stochastic version of the Ricker model in developing his synthetic likelihood.
- Motivation: hard to explore some parts of the parameter space in near chaotic models with low noise in the state model with full likelihood methods.
- Consider inference based on summary statistics for which the summary statistic likelihood is better behaved (Fasiolo, Pya and Wood, 2016).

Ricker model

Ricker, 1954, Wood, 2010

Ricker model summary statistics (Wood, 2010)

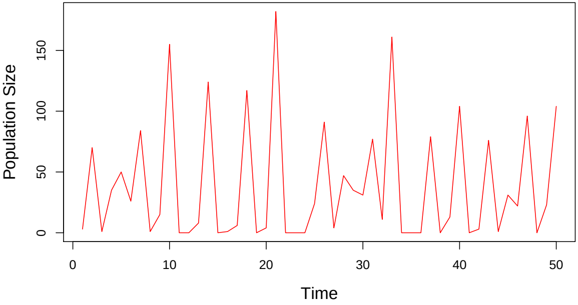
- Autocovariances to lag 5;
- The sample mean;
- Coefficients of cubic regression of ordered differences on observed values;
- Two coefficients of a certain autoregressive model;
- The number of zeros.

I have reduced these 13 summary statistics to 3 using the method of Fearnhead and Prangle (2012) (described later).

Ricker model

Ricker, 1954, Wood, 2010

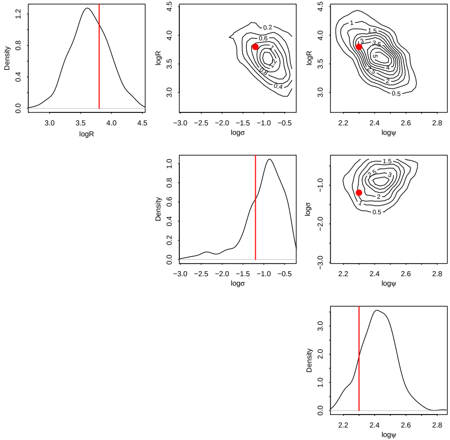
Simulated data: $(\log r, \log \sigma, \log \psi) = (3.8, -1.2, 2.3)$.



Motivating example: Ricker model

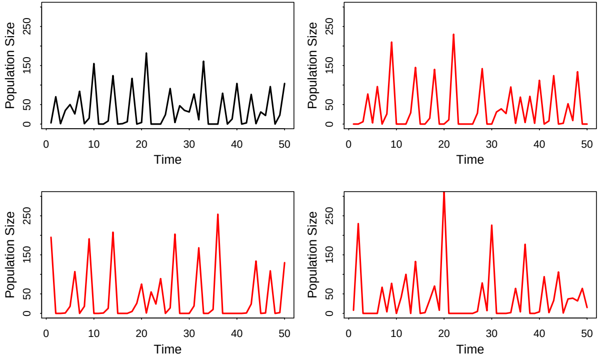
Ricker, 1954, Wood, 2010

Rejection ABC, keeping 500 samples from 100,000 prior samples, Euclidean distance for summaries scaled by prior predictive MAD



Ricker model

Ricker, 1954, Wood, 2010



Why should we believe an ABC analysis?

An ABC analysis approximates $p(\theta|y_{\text{obs}})$ or $p(\theta|S_{\text{obs}})$ but the approximation error is unknown. How can we validate the results?

- Posterior predictive checking of fitness for purpose (but if there is a problem is it the model or the computation?)
- Simple ABC methods which simulate from the prior allow repeated analyses using the same prior samples for different data - explore frequentist properties. They can also be used to check the reasonableness of the prior.
- Bayesian credible intervals have a coverage property under repeated sampling from the prior, check if ABC analyses are proper in this sense (Wegmann *et al.*, 2009, Prangle *et al.*, 2014).
- Compare results from different likelihood-free computational methods.
- Theoretical validations in an asymptotic sense.

Why should we believe an ABC analysis?

An ABC analysis approximates $p(\theta|y_{\text{obs}})$ or $p(\theta|S_{\text{obs}})$ but the approximation error is unknown. How can we validate the results?

- Posterior predictive checking of fitness for purpose (but if there is a problem is it the model or the computation?)
- Simple ABC methods which simulate from the prior allow repeated analyses using the same prior samples for different data - explore frequentist properties. They can also be used to check the reasonableness of the prior.
- Bayesian credible intervals have a coverage property under repeated sampling from the prior, check if ABC analyses are proper in this sense (Wegmann *et al.*, 2009, Prangle *et al.*, 2014).
- Compare results from different likelihood-free computational methods.
- Theoretical validations in an asymptotic sense.

Why should we believe an ABC analysis?

An ABC analysis approximates $p(\theta|y_{\text{obs}})$ or $p(\theta|S_{\text{obs}})$ but the approximation error is unknown. How can we validate the results?

- Posterior predictive checking of fitness for purpose (but if there is a problem is it the model or the computation?)
- Simple ABC methods which simulate from the prior allow repeated analyses using the same prior samples for different data - explore frequentist properties. They can also be used to check the reasonableness of the prior.
- Bayesian credible intervals have a coverage property under repeated sampling from the prior, check if ABC analyses are proper in this sense (Wegmann *et al.*, 2009, Prangle *et al.*, 2014).
- Compare results from different likelihood-free computational methods.
- Theoretical validations in an asymptotic sense.

Why should we believe an ABC analysis?

An ABC analysis approximates $p(\theta|y_{\text{obs}})$ or $p(\theta|S_{\text{obs}})$ but the approximation error is unknown. How can we validate the results?

- Posterior predictive checking of fitness for purpose (but if there is a problem is it the model or the computation?)
- Simple ABC methods which simulate from the prior allow repeated analyses using the same prior samples for different data - explore frequentist properties. They can also be used to check the reasonableness of the prior.
- Bayesian credible intervals have a coverage property under repeated sampling from the prior, check if ABC analyses are proper in this sense (Wegmann *et al.*, 2009, Prangle *et al.*, 2014).
- Compare results from different likelihood-free computational methods.
- Theoretical validations in an asymptotic sense.

Why should we believe an ABC analysis?

An ABC analysis approximates $p(\theta|y_{\text{obs}})$ or $p(\theta|S_{\text{obs}})$ but the approximation error is unknown. How can we validate the results?

- Posterior predictive checking of fitness for purpose (but if there is a problem is it the model or the computation?)
- Simple ABC methods which simulate from the prior allow repeated analyses using the same prior samples for different data - explore frequentist properties. They can also be used to check the reasonableness of the prior.
- Bayesian credible intervals have a coverage property under repeated sampling from the prior, check if ABC analyses are proper in this sense (Wegmann *et al.*, 2009, Prangle *et al.*, 2014).
- Compare results from different likelihood-free computational methods.
- Theoretical validations in an asymptotic sense.

Why should we believe an ABC analysis?

An ABC analysis approximates $p(\theta|y_{\text{obs}})$ or $p(\theta|S_{\text{obs}})$ but the approximation error is unknown. How can we validate the results?

- Posterior predictive checking of fitness for purpose (but if there is a problem is it the model or the computation?)
- Simple ABC methods which simulate from the prior allow repeated analyses using the same prior samples for different data - explore frequentist properties. They can also be used to check the reasonableness of the prior.
- Bayesian credible intervals have a coverage property under repeated sampling from the prior, check if ABC analyses are proper in this sense (Wegmann *et al.*, 2009, Prangle *et al.*, 2014).
- Compare results from different likelihood-free computational methods.
- Theoretical validations in an asymptotic sense.

Summary statistic choice

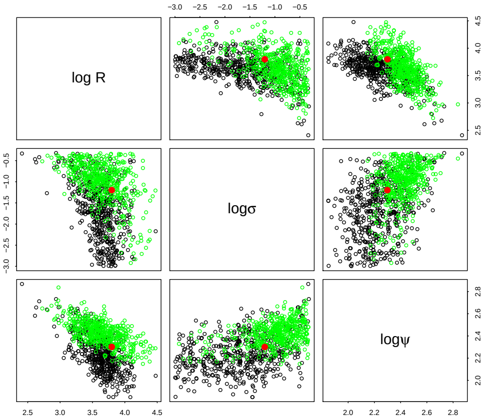
One of the most consequential assumptions made in an ABC analysis is:

$$p(\theta|y_{\text{obs}}) \approx p(\theta|S_{\text{obs}})$$

- More summary statistics do not necessarily equate to a better posterior approximation
  - High-dimensional summaries may make $p(\theta|y_{\text{obs}}) \approx p(\theta|S_{\text{obs}})$ more reasonable, but
  - High-dimensional summaries harder to match

Ricker model

ABC posterior samples: Fearnhead and Prangle summaries (green, 3 dimensional) and original summaries (black, 13 dimensional)



Summary statistic choice

- It may be difficult to automate summary statistic choice completely.
- The choice of summaries is sometimes motivated by concerns of model misspecification
  - A model for suitable $S = S(y)$ can be nearly well specified even if the model for y is not.
  - Matching summaries we care about makes sense.
- Modelling problems lead to computational problems.
- Difficult to simultaneously match inconsistent subsets of summaries (leading to a large ϵ for given computational effort).

Summary statistic choice

- It may be difficult to automate summary statistic choice completely.
- The choice of summaries is sometimes motivated by concerns of model misspecification
  - A model for suitable $S = S(y)$ can be nearly well specified even if the model for y is not.
  - Matching summaries we care about makes sense.
- Modelling problems lead to computational problems.
- Difficult to simultaneously match inconsistent subsets of summaries (leading to a large ϵ for given computational effort).

Summary statistic choice

- It may be difficult to automate summary statistic choice completely.
- The choice of summaries is sometimes motivated by concerns of model misspecification
  - A model for suitable $S = S(y)$ can be nearly well specified even if the model for y is not.
  - Matching summaries we care about makes sense.
- Modelling problems lead to computational problems.
- Difficult to simultaneously match inconsistent subsets of summaries (leading to a large ϵ for given computational effort).

Summary statistic choice

- It may be difficult to automate summary statistic choice completely.
- The choice of summaries is sometimes motivated by concerns of model misspecification
  - A model for suitable $S = S(y)$ can be nearly well specified even if the model for y is not.
  - Matching summaries we care about makes sense.
- Modelling problems lead to computational problems.
- Difficult to simultaneously match inconsistent subsets of summaries (leading to a large ϵ for given computational effort).

Summary statistic choice

- It may be difficult to automate summary statistic choice completely.
- The choice of summaries is sometimes motivated by concerns of model misspecification
  - A model for suitable $S = S(y)$ can be nearly well specified even if the model for y is not.
  - Matching summaries we care about makes sense.
- Modelling problems lead to computational problems.
- Difficult to simultaneously match inconsistent subsets of summaries (leading to a large ϵ for given computational effort).

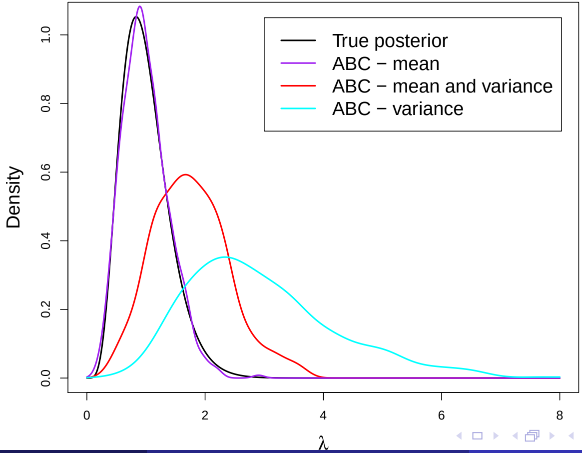
Summary statistic choice

- It may be difficult to automate summary statistic choice completely.
- The choice of summaries is sometimes motivated by concerns of model misspecification
  - A model for suitable $S = S(y)$ can be nearly well specified even if the model for y is not.
  - Matching summaries we care about makes sense.
- Modelling problems lead to computational problems.
- Difficult to simultaneously match inconsistent subsets of summaries (leading to a large ϵ for given computational effort).

Summary statistic choice

Sisson, Fan and Beaumont, 2018

$$y = (y_1, \ldots, y_5),\ y_i \sim \text{Poisson}(\lambda),\ \lambda \sim \text{Gamma}(1,1),$$
$$y_{\text{obs}} = (0,0,0,0,5),\ S = (\bar{y}, s^2),\ s_{\text{obs}} = (1,5).$$



Reducing summary statistic dimension

Blum *et al.*, 2013

Given a set of summary statistics, if it is beneficial to reduce dimension with minimal loss of statistical information (Blum *et al.*, 2013):

- Subset selection (Joyce and Marjoram, 2008, Nunes and Balding, 2010, Blum *et al.*, 2013).
- Projection methods such as partial least squares (Wegmann *et al.*, 2009), neural networks (Blum and François, 2010), decision theoretic approaches based on optimal point estimation (Fearnhead and Prangle, 2012).
- Regularization methods which use shrinkage estimation in conjunction with regression adjustments (Blum and François, 2010, Blum *et al.*, 2013).

Reducing summary statistic dimension

Blum *et al.*, 2013

Given a set of summary statistics, if it is beneficial to reduce dimension with minimal loss of statistical information (Blum *et al.*, 2013):

- Subset selection (Joyce and Marjoram, 2008, Nunes and Balding, 2010, Blum *et al.*, 2013).
- Projection methods such as partial least squares (Wegmann *et al.*, 2009), neural networks (Blum and François, 2010), decision theoretic approaches based on optimal point estimation (Fearnhead and Prangle, 2012).
- Regularization methods which use shrinkage estimation in conjunction with regression adjustments (Blum and François, 2010, Blum *et al.*, 2013).

Reducing summary statistic dimension
Blum *et al.*, 2013

Given a set of summary statistics, if it is beneficial to reduce dimension with minimal loss of statistical information (Blum *et al.*, 2013):

- Subset selection (Joyce and Marjoram, 2008, Nunes and Balding, 2010, Blum *et al.*, 2013).
- Projection methods such as partial least squares (Wegmann *et al.*, 2009), neural networks (Blum and François, 2010), decision theoretic approaches based on optimal point estimation (Fearnhead and Prangle, 2012).
- Regularization methods which use shrinkage estimation in conjunction with regression adjustments (Blum and François, 2010, Blum *et al.*, 2013).

Reducing summary statistic dimension

Given a set of summary statistics, if it is beneficial to reduce dimension with minimal loss of statistical information (Blum *et al.*, 2013):

- Subset selection (Joyce and Marjoram, 2008, Nunes and Balding, 2010, Blum *et al.*, 2013).
- Projection methods such as partial least squares (Wegmann *et al.*, 2009), neural networks (Blum and François, 2010), decision theoretic approaches based on optimal point estimation (Fearnhead and Prangle, 2012).
- Regularization methods which use shrinkage estimation in conjunction with regression adjustments (Blum and François, 2010, Blum *et al.*, 2013).

Steps of semi-automatic ABC:

- Define candidate summaries $u = (u_1, \ldots, u_C)$.
- For a pilot ABC run, determine a region of approximate posterior support.
- Simulate (θ, u) samples using the prior truncated to the training region of 2.
- For each component θ_j, $j = 1, \ldots, k$ of θ, fit a regression model, obtaining fitted values $\hat{\theta}_j(u)$.
- Use $\theta_j(u)$ as a reduced-dimension (one for each parameter) set of summary statistics.

Semi-automatic ABC

Fearnhead and Prangle, 2012

Steps of semi-automatic ABC:

- Define candidate summaries $u = (u_1, \ldots, u_C)$.
- For a pilot ABC run, determine a region of approximate posterior support.
- Simulate (θ, u) samples using the prior truncated to the training region of 2.
- For each component θ_j, $j = 1, \ldots, k$ of θ, fit a regression model, obtaining fitted values $\hat{\theta}_j(u)$.
- Use $\theta_j(u)$ as a reduced-dimension (one for each parameter) set of summary statistics.

Semi-automatic ABC

Fearnhead and Prangle, 2012

Steps of semi-automatic ABC:

- Define candidate summaries $u = (u_1, \ldots, u_C)$.
- For a pilot ABC run, determine a region of approximate posterior support.
- Simulate (θ, u) samples using the prior truncated to the training region of 2.
- For each component θ_j, $j = 1, \ldots, k$ of θ, fit a regression model, obtaining fitted values $\hat{\theta}_j(u)$.
- Use $\theta_j(u)$ as a reduced-dimension (one for each parameter) set of summary statistics.

Semi-automatic ABC

Fearnhead and Prangle, 2012

Steps of semi-automatic ABC:

- Define candidate summaries $u = (u_1, \ldots, u_C)$.
- For a pilot ABC run, determine a region of approximate posterior support.
- Simulate (θ, u) samples using the prior truncated to the training region of 2.
- For each component θ_j, $j = 1, \ldots, k$ of θ, fit a regression model, obtaining fitted values $\hat{\theta}_j(u)$.
- Use $\hat{\theta}_j(u)$ as a reduced-dimension (one for each parameter) set of summary statistics.

Semi-automatic ABC

Fearnhead and Prangle, 2012

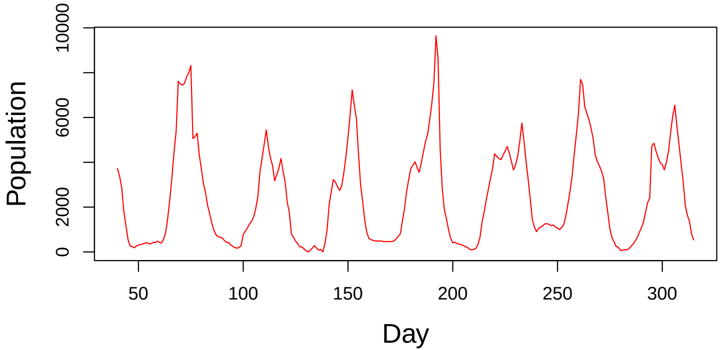
Steps of semi-automatic ABC:

- Define candidate summaries $u = (u_1, \ldots, u_C)$.
- For a pilot ABC run, determine a region of approximate posterior support.
- Simulate (θ, u) samples using the prior truncated to the training region of 2.
- For each component θ_j, $j = 1, \ldots, k$ of θ, fit a regression model, obtaining fitted values $\hat{\theta}_j(u)$.
- Use $\theta_j(u)$ as a reduced-dimension (one for each parameter) set of summary statistics.

Nicholson's Blowflies

Nicholson (1954, 1957)

Time series of blowfly population numbers (Nicholson, 1954, Figure 3)



Nicholson's Blowflies

Nicholson (1954, 1957)

Wood (2000) and Fasiolo and Wood (2016) discretize a delayed differential equation model (Gurney, Blythe and Nisbet, 1980):

For $t = 1, \ldots, T$,

$$
\begin{aligned}
n_t &= r_t + s_t, \\
r_t &\sim \text{Poisson}(P n_{t-\tau} \exp(-n_{t-\tau} e_t)) \\
s_t &\sim \text{Binomial}(n_{t-1}, \exp(-\delta \epsilon_t))
\end{aligned}
$$

where

- n_t is the population size at time t, r_t is a delayed recruitment process, s_t is an adult survival process.
- e_t, ϵ_t are gamma distributed noise sequences, mean 1 and respective standard deviations σ_p, σ_d

Nicholson's Blowflies

Nicholson (1954, 1957)

Summary statistics (Wood, 2010)

Twenty-three summary statisics, consisting of:

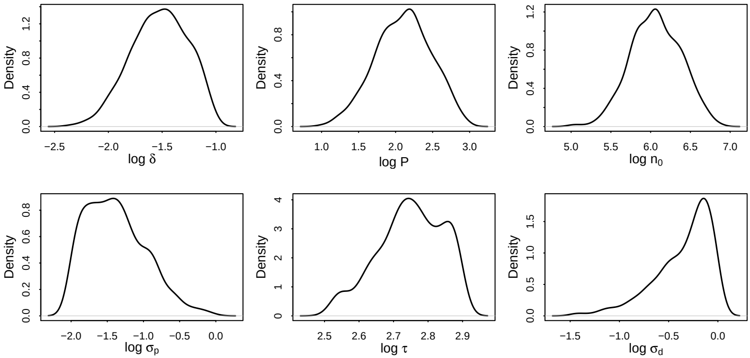
- Autocovariances to lag 11;
- The sample mean;
- The difference of sample mean and median;
- The number of observed turning points;
- Coefficients of cubic regression of ordered differences on observed values;
- Five coefficients of a certain autoregressive model;

Priors: $\delta \sim U[\exp(-3), \exp(-1)]$, $P \sim U[\exp(1), \exp(3)]$, $\log n_0 \sim U[\exp(5), \exp(7)]$, $\tau \sim U[\exp(2.5), \exp(2.9)]$, $\log \sigma_p \sim U[-2, 0]$, $\log \sigma_d \sim U[-1.5, 0]$.

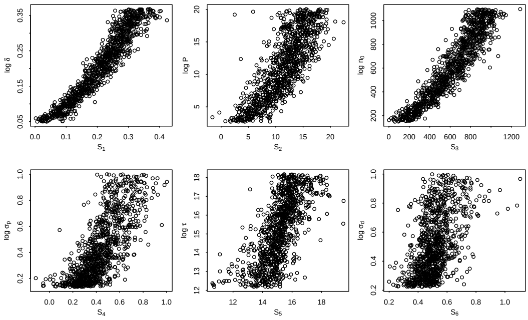
Nicholson's Blowflies

Nicholson (1954, 1957)

Rejection ABC, Fearnhead and Prangle summaries, 100,000 prior samples, ϵ retaining 500 samples.



Fearnhead and Prangle summary statistics



Further refinements - regression

Beaumont, Zhang and Balding, 2002, Blum, 2010, Blum and François, 2010

- Consider $(\theta_i, S_i) \sim p(\theta)p(S|\theta)$, $i = 1, \ldots, N$ and the regression model

$$\theta_i = \beta_0 + \beta^T(S_i - S_{\text{obs}}) + \eta_i$$

where η_i are mean zero iid.

- Empirical residuals, $\hat{\eta}_i = \theta_i - \hat{\beta}_0 - \hat{\beta}^T(S_i - S_{\text{obs}})$.

Regression adjusted samples:

Fitted mean at S_{obs} (i.e. $\hat{\beta}_0$) plus empirical residuals:

$$\theta_i^a = \hat{\beta}_0 + \hat{\eta}_i = \theta_i - \hat{\beta}^T(S_i - S_{\text{obs}}), \quad i = 1, \ldots, N$$

- Fitting can be localized, weighting, multivariate θ.

Further refinements - regression

Beaumont, Zhang and Balding, 2002, Blum, 2010, Blum and François, 2010

- Consider $(\theta_i, S_i) \sim p(\theta)p(S|\theta)$, $i = 1, \ldots, N$ and the regression model

$$\theta_i = \beta_0 + \beta^T(S_i - S_{\text{obs}}) + \eta_i$$

where η_i are mean zero iid.

- Empirical residuals, $\hat{\eta}_i = \theta_i - \hat{\beta}_0 - \hat{\beta}^T(S_i - S_{\text{obs}})$.

Regression adjusted samples:

Fitted mean at S_{obs} (i.e. $\hat{\beta}_0$) plus empirical residuals:

$$\theta_i^a = \hat{\beta}_0 + \hat{\eta}_i = \theta_i - \hat{\beta}^T(S_i - S_{\text{obs}}), \quad i = 1, \ldots, N$$

- Fitting can be localized, weighting, multivariate θ.

Further refinements - regression

Beaumont, Zhang and Balding, 2002, Blum, 2010, Blum and François, 2010

- Consider $(\theta_i, S_i) \sim p(\theta)p(S|\theta)$, $i = 1, \ldots, N$ and the regression model

$$\theta_i = \beta_0 + \beta^T(S_i - S_{\text{obs}}) + \eta_i$$

  where η_i are mean zero iid.
- Empirical residuals, $\hat{\eta}_i = \theta_i - \hat{\beta}_0 - \hat{\beta}^T(S_i - S_{\text{obs}})$.

Regression adjusted samples:

Fitted mean at S_{obs} (i.e. $\hat{\beta}_0$) plus empirical residuals:

$$\theta_i^a = \hat{\beta}_0 + \hat{\eta}_i = \theta_i - \hat{\beta}^T(S_i - S_{\text{obs}}), \quad i = 1, \ldots, N$$

- Fitting can be localized, weighting, multivariate θ.

Further refinements - regression

Beaumont, Zhang and Balding, 2002, Blum, 2010, Blum and François, 2010

- Consider $(\theta_i, S_i) \sim p(\theta)p(S|\theta)$, $i = 1, \ldots, N$ and the regression model

$$\theta_i = \beta_0 + \beta^T(S_i - S_{\text{obs}}) + \eta_i$$

  where η_i are mean zero iid.
- Empirical residuals, $\hat{\eta}_i = \theta_i - \hat{\beta}_0 - \hat{\beta}^T(S_i - S_{\text{obs}})$.

Regression adjusted samples:

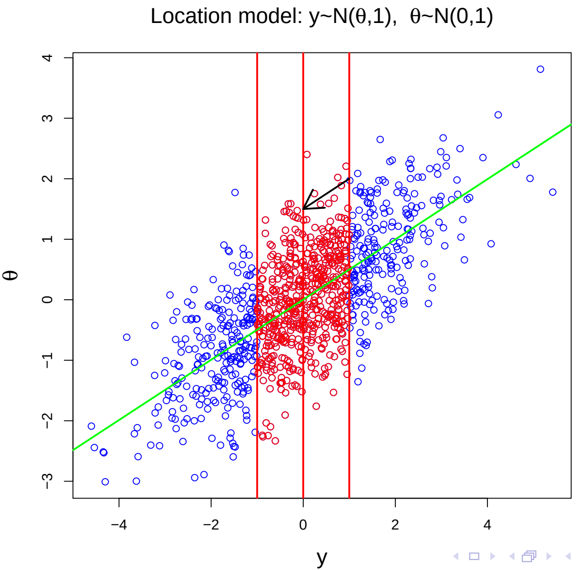
Fitted mean at S_{obs} (i.e. $\hat{\beta}_0$) plus empirical residuals:

$$\theta_i^a = \hat{\beta}_0 + \hat{\eta}_i = \theta_i - \hat{\beta}^T(S_i - S_{\text{obs}}), \quad i = 1, \ldots, N$$

- Fitting can be localized, weighting, multivariate θ.

Further refinements - regression

Beaumont, Zhang and Balding, 2002, Blum, 2010, Blum and François, 2010



Further refinements - regression

Beaumont, Zhang and Balding, 2002, Blum, 2010, Blum and François, 2010

- A nonlinear regression adjustment considers the model

$$\theta_i = \mu(S_i) + \sigma(S_i)\eta_i,$$

where the η_i are mean zero variance one.

- With estimates $\hat{\mu}(\cdot)$ and $\hat{\sigma}(\cdot)$ of $\mu(\cdot)$ and $\sigma(\cdot)$, standardized empirical residuals are

$$\hat{\eta}_i = \frac{\theta_i - \hat{\mu}(S_i)}{\hat{\sigma}(S_i)}.$$

- Adjusted posterior samples (using the fitted regression and empirical residuals):

$$\theta_i^a = \hat{\mu}(S_{\text{obs}}) + \hat{\sigma}(S_{\text{obs}})\frac{\theta_i - \hat{\mu}(S_i)}{\hat{\sigma}(S_i)}.$$

- A nonlinear regression adjustment considers the model

$$\theta_i = \mu(S_i) + \sigma(S_i)\eta_i,$$

where the η_i are mean zero variance one.

- With estimates $\hat{\mu}(\cdot)$ and $\hat{\sigma}(\cdot)$ of $\mu(\cdot)$ and $\sigma(\cdot)$, standardized empirical residuals are

$$\hat{\eta}_i = \frac{\theta_i - \hat{\mu}(S_i)}{\hat{\sigma}(S_i)}.$$

- Adjusted posterior samples (using the fitted regression and empirical residuals):

$$\theta_i^a = \hat{\mu}(S_{\text{obs}}) + \hat{\sigma}(S_{\text{obs}})\frac{\theta_i - \hat{\mu}(S_i)}{\hat{\sigma}(S_i)}.$$

Further refinements - regression

Beaumont, Zhang and Balding, 2002, Blum, 2010, Blum and François, 2010

- A nonlinear regression adjustment considers the model

$$\theta_i = \mu(S_i) + \sigma(S_i)\eta_i,$$

where the η_i are mean zero variance one.

- With estimates $\hat{\mu}(\cdot)$ and $\hat{\sigma}(\cdot)$ of $\mu(\cdot)$ and $\sigma(\cdot)$, standardized empirical residuals are

$$\hat{\eta}_i = \frac{\theta_i - \hat{\mu}(S_i)}{\hat{\sigma}(S_i)}.$$

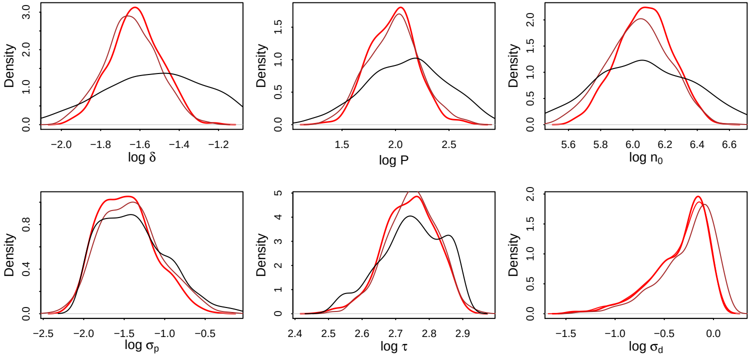
- Adjusted posterior samples (using the fitted regression and empirical residuals):

$$\theta_i^a = \hat{\mu}(S_{\text{obs}}) + \hat{\sigma}(S_{\text{obs}})\frac{\theta_i - \hat{\mu}(S_i)}{\hat{\sigma}(S_i)}.$$

Nicholson's Blowflies

Nicholson (1954, 1957)

Neural network regression adjustment (red), local linear (brown), none (black).



Generalizing simple rejection ABC

Simple rejection ABC simulates (θ, S) from

$$p_\epsilon(\theta, S|S_{obs}) \propto p(\theta)p(S|\theta)I(d(S, S_{obs}) < \epsilon),$$

with the θ marginal

$$p_\epsilon(\theta|S_{obs}) \propto p(\theta) \int I(d(S, S_{obs}) < \epsilon)p(S|\theta)dS.$$

Hence instead of $p(S_{\mathrm{obs}}|\theta)$ we are using the approximate likelihood:

$$p_\epsilon(S_{\mathrm{obs}}|\theta) \propto \int I(d(S, S_{obs}) < \epsilon)p(S|\theta)dS.$$

Generalizing simple rejection ABC

Simple rejection ABC simulates (θ, S) from

$$p_\epsilon(\theta, S|S_{obs}) \propto p(\theta)p(S|\theta)I(d(S, S_{obs}) < \epsilon),$$

with the θ marginal

$$p_\epsilon(\theta|S_{obs}) \propto p(\theta) \int I(d(S, S_{obs}) < \epsilon)p(S|\theta)dS.$$

Hence instead of $p(S_{\text{obs}}|\theta)$ we are using the approximate likelihood:

$$p_\epsilon(S_{\text{obs}}|\theta) \propto \int I(d(S, S_{obs}) < \epsilon)p(S|\theta)dS.$$

Generalizing simple rejection ABC

For a "nice" kernel function $K(u)$ and $K_\epsilon(u) = \epsilon^{-1}K(\epsilon^{-1}u)$ where $\epsilon > 0$ is a bandwidth we can consider instead

Kernel ABC likelihood

$$p_{\epsilon,K}(S_{\text{obs}}|\theta) = \int K_\epsilon(d(S, S_{obs}))p(S|\theta)dS,$$

which approaches $p(S_{obs}|\theta)$ as $\epsilon \to 0$.

Generalizing simple rejection ABC

The posterior approximation

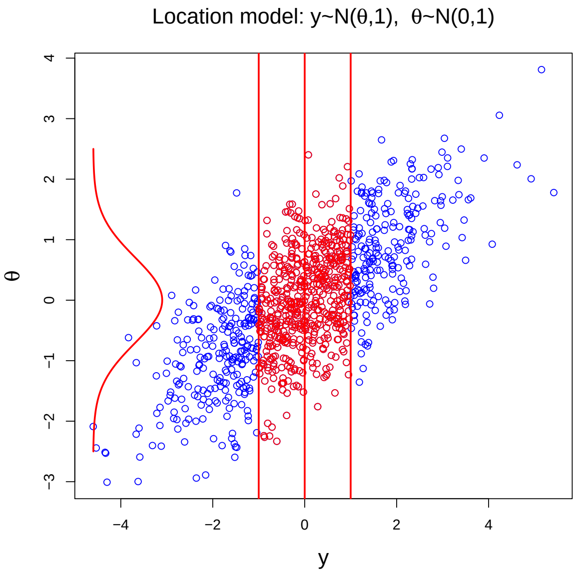
$$p_{\epsilon,K}(\theta|S_{obs}) \propto p(\theta)p_{\epsilon,K}(S_{\text{obs}}|\theta) = p(\theta)\int K_\epsilon(d(S,S_{obs})p(S|\theta)dS,$$

is the θ marginal of

$$p_{\epsilon,K}(\theta,S|S_{obs}) \propto p(\theta)p(S|\theta)K_\epsilon(d(S,S_{obs})),$$

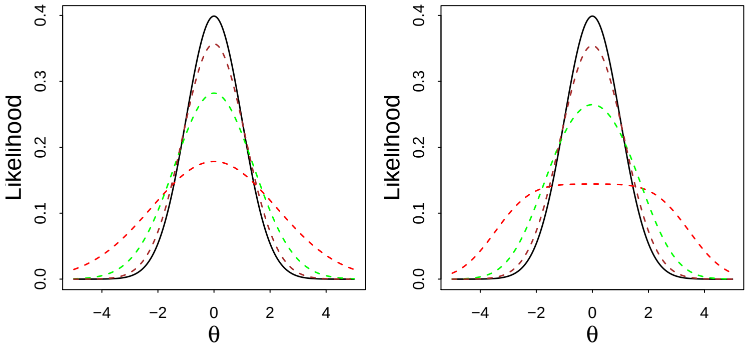
and the basic rejection algorithm can be generalized.

Generalizing simple rejection ABC



Genearlizing simple rejection ABC

Kernel ABC likelihoods for toy example, Gaussian (left) and uniform (right) kernels, $\epsilon = 2, 1, 0.5$ (red, green, brown) Black is the truth.



Importance sampling

For some density function $f(\theta)$ and suitable functions $h(\theta)$ we want to approximate expectations

$$E_f(h(\theta)) = \int h(\theta)f(\theta)d\theta.$$

$f(\theta)$ might be a posterior distribution and $E_f(h(\theta))$ a posterior expectation.

Let $g(\theta)$ be an (importance) density, easy to sample, then

$$E_f(h(\theta)) = \int h(\theta)f(\theta)d\theta = \int h(\theta)\frac{f(\theta)}{g(\theta)}g(\theta)d\theta = E_g\left(h(\theta)\frac{f(\theta)}{g(\theta)}\right),$$

an expectation with respect to $g(\theta)$.

Importance sampling

For some density function $f(\theta)$ and suitable functions $h(\theta)$ we want to approximate expectations

$$E_f(h(\theta)) = \int h(\theta)f(\theta)d\theta.$$

$f(\theta)$ might be a posterior distribution and $E_f(h(\theta))$ a posterior expectation.

Let $g(\theta)$ be an (importance) density, easy to sample, then

$$E_f(h(\theta)) = \int h(\theta)f(\theta)d\theta = \int h(\theta)\frac{f(\theta)}{g(\theta)}g(\theta)d\theta = E_g\left(h(\theta)\frac{f(\theta)}{g(\theta)}\right),$$

an expectation with respect to $g(\theta)$.

Importance sampling

For some density function $f(\theta)$ and suitable functions $h(\theta)$ we want to approximate expectations

$$E_f(h(\theta)) = \int h(\theta)f(\theta)d\theta.$$

$f(\theta)$ might be a posterior distribution and $E_f(h(\theta))$ a posterior expectation.

Let $g(\theta)$ be an (importance) density, easy to sample, then

$$E_f(h(\theta)) = \int h(\theta)f(\theta)d\theta = \int h(\theta)\frac{f(\theta)}{g(\theta)}g(\theta)d\theta = E_g\left(h(\theta)\frac{f(\theta)}{g(\theta)}\right),$$

an expectation with respect to $g(\theta)$.

Importance sampling

We can approximate

$$E_f(h(\theta)) = E_g\left(h(\theta)\frac{f(\theta)}{g(\theta)}\right),$$

by

$$\sum_{j=1}^{J} w^j h(\theta^{(j)}),$$

where $\theta^{(1)},\ldots,\theta^{(J)} \sim g(\theta)$, and $w^j = (1/J)f(\theta^{(j)})/g(\theta^{(j)})$.

The choice of importance density is important. If $f(\theta)$ is unnormalized, scale the weights so that $\sum_j w^j = 1$.

To sample from

$$p_{\epsilon,K}(\theta, S|S_{obs}) \propto p(\theta)p(S|\theta)K_\epsilon(d(S, S_{obs})),$$

consider an importance density $g(\theta)p(S|\theta)$.

The intractable likelihood $p(S|\theta)$ cancels out in the importance weights,

$$w^j \propto \frac{p(\theta^{(j)})p(S|\theta^{(j)})K_\epsilon(d(S, S_{obs}))}{g(\theta^{(j)})p(S|\theta^{(j)})} = \frac{p(\theta^{(j)})K_\epsilon(d(S, S_{obs}))}{g(\theta^{(j)})}.$$

With a compact kernel we may choose ϵ so that a certain fraction of prior samples receive positive weight.

Importance sampling ABC

To sample from

$$p_{\epsilon,K}(\theta, S|S_{obs}) \propto p(\theta)p(S|\theta)K_\epsilon(d(S, S_{obs})),$$

consider an importance density $g(\theta)p(S|\theta)$.

The intractable likelihood $p(S|\theta)$ cancels out in the importance weights,

$$w^j \propto \frac{p(\theta^{(j)})p(S|\theta^{(j)})K_\epsilon(d(S, S_{obs}))}{g(\theta^{(j)})p(S|\theta^{(j)})} = \frac{p(\theta^{(j)})K_\epsilon(d(S, S_{obs}))}{g(\theta^{(j)})}.$$

With a compact kernel we may choose ϵ so that a certain fraction of prior samples receive positive weight.

Importance sampling ABC

To sample from

$$p_{\epsilon,K}(\theta, S|S_{obs}) \propto p(\theta)p(S|\theta)K_\epsilon(d(S, S_{obs})),$$

consider an importance density $g(\theta)p(S|\theta)$.

The intractable likelihood $p(S|\theta)$ cancels out in the importance weights,

$$w^j \propto \frac{p(\theta^{(j)})p(S|\theta^{(j)})K_\epsilon(d(S, S_{obs}))}{g(\theta^{(j)})p(S|\theta^{(j)})} = \frac{p(\theta^{(j)})K_\epsilon(d(S, S_{obs}))}{g(\theta^{(j)})}.$$

With a compact kernel we may choose ϵ so that a certain fraction of prior samples receive positive weight.

Markov chain Monte Carlo (MCMC)

MCMC simulates a Markov chain $\{\theta^{(n)}; n = 0, 1, \dots\}$ on the parameter space such that

- The Markov chain has stationary distribution $p(\theta|y)$
- The Markov chain can easily be simulated

Run the chain, keeping samples after convergence to get a dependent sample from the posterior.

Metropolis-Hastings algorithm

Metropolis *et al.*, 1953, Hastings, 1970

Metropolis-Hastings algorithm

For current value θ generate a proposal θ' from $q(\theta'|\theta)$, accepting with probability

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

$$= \min\left\{1, \text{Prior ratio} \times \text{Likelihood ratio} \times \text{Proposal ratio}\right\},$$

with the current value retained otherwise.

ABC-MCMC

Marjoram *et al.*, 2003

Consider a Metropolis-Hastings algorithm for the target ABC posterior

$$p_{\epsilon,K}(\theta, S|S_{obs}) \propto p(\theta)p(S|\theta)K_\epsilon(d(S, S_{obs})).$$

Letting the proposal density be $g(\theta'|\theta)p(S'|\theta')$, The intractable terms cancel from the likelihood and proposal ratios in the acceptance probability:

$$\min\left\{1, \frac{p(\theta')p(S'|\theta')K_\epsilon(d(S', S_{obs}))}{p(\theta)p(S|\theta)K_\epsilon(d(S, S_{obs}))}\frac{g(\theta|\theta')p(S|\theta)}{g(\theta'|\theta)p(S'|\theta')}\right\}$$

$$= \min\left\{1, \frac{p(\theta')K_\epsilon(d(S', S_{obs}))}{p(\theta)K_\epsilon(d(S, S_{obs}))}\frac{g(\theta|\theta')}{g(\theta'|\theta)}\right\}.$$

ABC-MCMC

Marjoram *et al.*, 2003

Consider a Metropolis-Hastings algorithm for the target ABC posterior

$$p_{\epsilon,K}(\theta, S|S_{obs}) \propto p(\theta)p(S|\theta)K_\epsilon(d(S, S_{obs})).$$

Letting the proposal density be $g(\theta'|\theta)p(S'|\theta')$, The intractable terms cancel from the likelihood and proposal ratios in the acceptance probability:

$$\min\left\{1, \frac{p(\theta')p(S'|\theta')K_\epsilon(d(S', S_{obs}))}{p(\theta)p(S|\theta)K_\epsilon(d(S, S_{obs}))}\frac{g(\theta|\theta')p(S|\theta)}{g(\theta'|\theta)p(S'|\theta')}\right\}$$

$$= \min\left\{1, \frac{p(\theta')K_\epsilon(d(S', S_{obs}))}{p(\theta)K_\epsilon(d(S, S_{obs}))}\frac{g(\theta|\theta')}{g(\theta'|\theta)}\right\}.$$

Markov chain Monte Carlo

Marjoram *et al.*, 2003

Difficulties with the basic ABC-MCMC

- It can be hard to move if we are in the tails of the posterior, as summary statistics may be hard to match there.
- It may be hard to calibrate the tolerance (i.e. the value of ϵ). The tolerance determines the mixing rate as well as the accuracy.

More advanced ABC-MCMC methods are available that mitigate these problems to some extent.

Markov chain Monte Carlo

Marjoram *et al.*, 2003

Difficulties with the basic ABC-MCMC

- It can be hard to move if we are in the tails of the posterior, as summary statistics may be hard to match there.
- It may be hard to calibrate the tolerance (i.e. the value of ϵ). The tolerance determines the mixing rate as well as the accuracy.

More advanced ABC-MCMC methods are available that mitigate these problems to some extent.

Wegmann, Leuenberger and Excoffier (2009) suggest

- Applying ABC-MCMC with a fairly large tolerance to ensure good mixing;
- Extracting a subset of samples for which the simulated summaries are closes to the observed values;
- Applying postprocessing regression adjustments;
- They also suggest partial least squares dimension reductions for summaries

Fewer simulations are required compared to simple rejection ABC. Bortot, Coles and Sisson (2007) incorporate ϵ into the target distribution. Baragatti, Grimaud and Pommeret (2013) consider parallel tempering.

Wegmann, Leuenberger and Excoffier (2009) suggest

- Applying ABC-MCMC with a fairly large tolerance to ensure good mixing;
- Extracting a subset of samples for which the simulated summaries are closes to the observed values;
- Applying postprocessing regression adjustments;
- They also suggest partial least squares dimension reductions for summaries

Fewer simulations are required compared to simple rejection ABC. Bortot, Coles and Sisson (2007) incorporate ϵ into the target distribution. Baragatti, Grimaud and Pommeret (2013) consider parallel tempering.

Markov chain Monte Carlo ABC

Wegmann, Leuenberger and Excoffier, 2009

Wegmann, Leuenberger and Excoffier (2009) suggest

- Applying ABC-MCMC with a fairly large tolerance to ensure good mixing;
- Extracting a subset of samples for which the simulated summaries are closes to the observed values;
- Applying postprocessing regression adjustments;
- They also suggest partial least squares dimension reductions for summaries

Fewer simulations are required compared to simple rejection ABC. Bortot, Coles and Sisson (2007) incorporate ϵ into the target distribution. Baragatti, Grimaud and Pommeret (2013) consider parallel tempering.

Wegmann, Leuenberger and Excoffier (2009) suggest

- Applying ABC-MCMC with a fairly large tolerance to ensure good mixing;
- Extracting a subset of samples for which the simulated summaries are closes to the observed values;
- Applying postprocessing regression adjustments;
- They also suggest partial least squares dimension reductions for summaries

Fewer simulations are required compared to simple rejection ABC. Bortot, Coles and Sisson (2007) incorporate ϵ into the target distribution. Baragatti, Grimaud and Pommeret (2013) consider parallel tempering.

Markov chain Monte Carlo ABC

Wegmann, Leuenberger and Excoffier, 2009

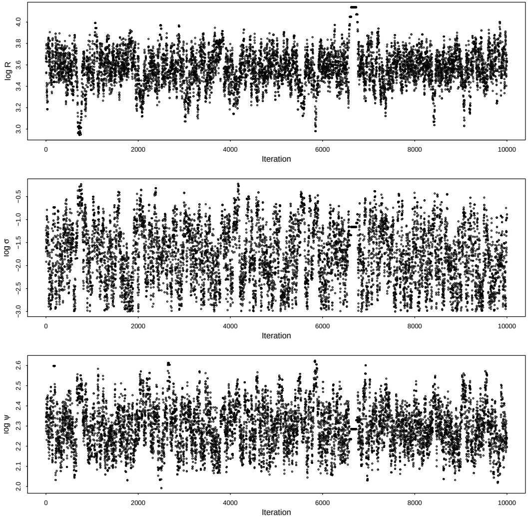
Wegmann, Leuenberger and Excoffier (2009) suggest

- Applying ABC-MCMC with a fairly large tolerance to ensure good mixing;
- Extracting a subset of samples for which the simulated summaries are closes to the observed values;
- Applying postprocessing regression adjustments;
- They also suggest partial least squares dimension reductions for summaries

Fewer simulations are required compared to simple rejection ABC. Bortot, Coles and Sisson (2007) incorporate ϵ into the target distribution. Baragatti, Grimaud and Pommeret (2013) consider parallel tempering.

Markov chain Monte Carlo ABC

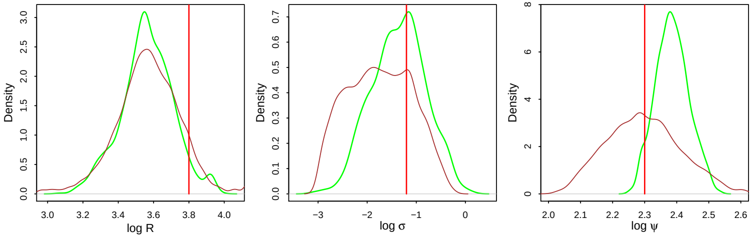
Wegmann, Leuenberger and Excoffier, 2009



Markov chain Monte Carlo ABC

Wegmann, Leuenberger and Excoffier, 2009

Density estimates using the raw MCMC output (brown) and after subsetting and regression (green)



The pseudo-marginal perspective

Beaumont, 2003, Andrieu and Roberts, 2009

ABC-MCMC algorithms are examples of pseudo-marginal Metropolis-Hastings algorithms.

Suppose we have a parameter θ and data y_{obs} and consider a Metropolis-Hastings algorithm with proposal $g(\theta'|\theta)$.

Replace $p(y_{\text{obs}}|\theta)$ by a non-negative unbiased estimate $\hat{p}(y_{\text{obs}}|\theta)$ of it in the acceptance probability: acceptance probability

$$\min\left\{1, \frac{p(\theta')\hat{p}(y_{\text{obs}}|\theta')}{p(\theta)\hat{p}(y_{\text{obs}}|\theta)}\frac{g(\theta|\theta')}{g(\theta'|\theta)}\right\}.$$

Surprisingly, this modified algorithm is exact.

The pseudo-marginal perspective

Beaumont, 2003, Andrieu and Roberts, 2009

ABC-MCMC algorithms are examples of pseudo-marginal Metropolis-Hastings algorithms.

Suppose we have a parameter θ and data y_{obs} and consider a Metropolis-Hastings algorithm with proposal $g(\theta'|\theta)$.

Replace $p(y_{\text{obs}}|\theta)$ by a non-negative unbiased estimate $\hat{p}(y_{\text{obs}}|\theta)$ of it in the acceptance probability: acceptance probability

$$\min\left\{1, \frac{p(\theta')\hat{p}(y_{\text{obs}}|\theta')}{p(\theta)\hat{p}(y_{\text{obs}}|\theta)}\frac{g(\theta|\theta')}{g(\theta'|\theta)}\right\}.$$

Surprisingly, this modified algorithm is exact.

The pseudo-marginal perspective

Beaumont, 2003, Andrieu and Roberts, 2009

ABC-MCMC algorithms are examples of pseudo-marginal Metropolis-Hastings algorithms.

Suppose we have a parameter θ and data y_{obs} and consider a Metropolis-Hastings algorithm with proposal $g(\theta'|\theta)$.

Replace $p(y_{\text{obs}}|\theta)$ by a non-negative unbiased estimate $\hat{p}(y_{\text{obs}}|\theta)$ of it in the acceptance probability: acceptance probability

$$\min\left\{1, \frac{p(\theta')\hat{p}(y_{\text{obs}}|\theta')}{p(\theta)\hat{p}(y_{\text{obs}}|\theta)}\frac{g(\theta|\theta')}{g(\theta'|\theta)}\right\}.$$

Surprisingly, this modified algorithm is exact.

The pseudo-marginal perspective

Beaumont, 2003, Andrieu and Roberts, 2009

ABC-MCMC algorithms are examples of pseudo-marginal Metropolis-Hastings algorithms.

Suppose we have a parameter θ and data y_{obs} and consider a Metropolis-Hastings algorithm with proposal $g(\theta'|\theta)$.

Replace $p(y_{\text{obs}}|\theta)$ by a non-negative unbiased estimate $\hat{p}(y_{\text{obs}}|\theta)$ of it in the acceptance probability: acceptance probability

$$\min\left\{1, \frac{p(\theta')\hat{p}(y_{\text{obs}}|\theta')}{p(\theta)\hat{p}(y_{\text{obs}}|\theta)} \frac{g(\theta|\theta')}{g(\theta'|\theta)}\right\}.$$

Surprisingly, this modified algorithm is exact.

The pseudo-marginal perspective

Beaumont, 2003, Andrieu and Roberts, 2009

Consider once again the likelihood approximation

$$p_{\epsilon,K}(S_{obs}|\theta) = \int K_\epsilon(d(S, S_{obs}))p(S|\theta)dS.$$

A non-negative unbiased estimate of this likelihood is

$$K_\epsilon(d(S', S_{obs})) \quad S' \sim p(S|\theta)$$

We can average over more than one draw in obtaining the likelihood estimate but one draw is recommended (Bornn *et al.*, 2017).

The pseudo-marginal perspective

Beaumont, 2003, Andrieu and Roberts, 2009

Consider once again the likelihood approximation

$$p_{\epsilon,K}(S_{obs}|\theta) = \int K_\epsilon(d(S, S_{obs}))p(S|\theta)dS.$$

A non-negative unbiased estimate of this likelihood is

$$K_\epsilon(d(S', S_{obs})) \quad S' \sim p(S|\theta)$$

We can average over more than one draw in obtaining the likelihood estimate but one draw is recommended (Bornn *et al.*, 2017).

The pseudo-marginal perspective

Beaumont, 2003, Andrieu and Roberts, 2009

Consider once again the likelihood approximation

$$p_{\epsilon,K}(S_{obs}|\theta) = \int K_\epsilon(d(S, S_{obs}))p(S|\theta)dS.$$

A non-negative unbiased estimate of this likelihood is

$$K_\epsilon(d(S', S_{obs})) \quad S' \sim p(S|\theta)$$

We can average over more than one draw in obtaining the likelihood estimate but one draw is recommended (Bornn *et al.*, 2017).

The pseudo-marginal perspective

Beaumont, 2003, Andrieu and Roberts, 2009

Examining the ABC-MCMC acceptance probability

$$\min\left\{1, \frac{p(\theta')K_\epsilon(d(S', S_{obs}))}{p(\theta)K_\epsilon(d(S, S_{obs}))}\frac{g(\theta|\theta')}{g(\theta'|\theta)}\right\}$$

we see that ABC-MCMC is just a pseudo-marginal algorithm for sampling $p_{\epsilon,K}(\theta|S_{obs})$.

SMC samplers

An alternative to MCMC methods is to use sequential Monte Carlo samplers (Del Moral *et al.*, 2006).

Advantages

- Easy adaptive design of proposals;
- Better performance for irregular (for example multi-modal) target distributions.

A sequence of target distributions $p_{\epsilon_i,K}(\theta|S_{obs})$ are considered for tolerances $\epsilon_1 > \cdots > \epsilon_T$.

A population of weighted particles is maintained as the tolerances are traversed sequentially, starting with the largest tolerance (Sisson *et al.*, 2007, Beaumont *et al.*, 2009, Toni *et al.*, 2009, Drovandi and Pettit, 2011, Del Moral *et al.*, 2012).

SMC samplers

An alternative to MCMC methods is to use sequential Monte Carlo samplers (Del Moral *et al.*, 2006).

Advantages

- Easy adaptive design of proposals;
- Better performance for irregular (for example multi-modal) target distributions.

A sequence of target distributions $p_{\epsilon_i,K}(\theta|S_{obs})$ are considered for tolerances $\epsilon_1 > \cdots > \epsilon_T$.

A population of weighted particles is maintained as the tolerances are traversed sequentially, starting with the largest tolerance (Sisson *et al.*, 2007, Beaumont *et al.*, 2009, Toni *et al.*, 2009, Drovandi and Pettit, 2011, Del Moral *et al.*, 2012).

SMC samplers

An alternative to MCMC methods is to use sequential Monte Carlo samplers (Del Moral *et al.*, 2006).

Advantages

- Easy adaptive design of proposals;
- Better performance for irregular (for example multi-modal) target distributions.

A sequence of target distributions $p_{\epsilon_i,K}(\theta|S_{obs})$ are considered for tolerances $\epsilon_1 > \cdots > \epsilon_T$.

A population of weighted particles is maintained as the tolerances are traversed sequentially, starting with the largest tolerance (Sisson *et al.*, 2007, Beaumont *et al.*, 2009, Toni *et al.*, 2009, Drovandi and Pettit, 2011, Del Moral *et al.*, 2012).

Synthetic likelihood

Wood, 2010

- There is a curse of dimensionality that is inherent in ABC methods.
  - The ABC likelihood is in effect based on a kernel estimation of the summary statistic distribution.
  - When the number of summary statistics is large, this becomes impractical.
- An alternative likelihood free methodology is synthetic likelihood.

Synthetic likelihood

At each θ, simulate $S_1, \ldots, S_N \sim p(S|\theta)$,

$$\hat{\mu}(\theta) = \frac{1}{N}\sum_{i=1}^N S_i, \quad \hat{\Sigma}(\theta) = \frac{1}{N-1}\sum_{i=1}^N (S_i - \hat{\mu}(\theta))(S_i - \hat{\mu}(\theta))^T$$

and use as the likelihood $\phi(S_{obs}; \hat{\mu}(\theta), \hat{\Sigma}(\theta))$ where $\phi(z; \mu, \Sigma)$ is the multivariate normal density with mean μ and covariance Σ.

Synthetic likelihood

Wood, 2010

- There is a curse of dimensionality that is inherent in ABC methods.
  - The ABC likelihood is in effect based on a kernel estimation of the summary statistic distribution.
  - When the number of summary statistics is large, this becomes impractical.
- An alternative likelihood free methodology is synthetic likelihood.

Synthetic likelihood

At each θ, simulate $S_1, \ldots, S_N \sim p(S|\theta)$,

$$\hat{\mu}(\theta) = \frac{1}{N}\sum_{i=1}^{N} S_i, \quad \hat{\Sigma}(\theta) = \frac{1}{N-1}\sum_{i=1}^{N}(S_i - \hat{\mu}(\theta))(S_i - \hat{\mu}(\theta))^T$$

and use as the likelihood $\phi(S_{obs}; \hat{\mu}(\theta), \hat{\Sigma}(\theta))$ where $\phi(z; \mu, \Sigma)$ is the multivariate normal density with mean μ and covariance Σ.

Synthetic likelihood

Wood, 2010

- There is a curse of dimensionality that is inherent in ABC methods.
  - The ABC likelihood is in effect based on a kernel estimation of the summary statistic distribution.
  - When the number of summary statistics is large, this becomes impractical.
- An alternative likelihood free methodology is synthetic likelihood.

Synthetic likelihood

At each θ, simulate $S_1, \ldots, S_N \sim p(S|\theta)$,

$$\hat{\mu}(\theta) = \frac{1}{N} \sum_{i=1}^{N} S_i, \quad \hat{\Sigma}(\theta) = \frac{1}{N-1} \sum_{i=1}^{N} (S_i - \hat{\mu}(\theta))(S_i - \hat{\mu}(\theta))^T$$

and use as the likelihood $\phi(S_{obs}; \hat{\mu}(\theta), \hat{\Sigma}(\theta))$ where $\phi(z; \mu, \Sigma)$ is the multivariate normal density with mean μ and covariance Σ.

Synthetic likelihood

Wood, 2010

- There is a curse of dimensionality that is inherent in ABC methods.
  - The ABC likelihood is in effect based on a kernel estimation of the summary statistic distribution.
  - When the number of summary statistics is large, this becomes impractical.
- An alternative likelihood free methodology is synthetic likelihood.

Synthetic likelihood

At each θ, simulate $S_1, \ldots, S_N \sim p(S|\theta)$,

$$\hat{\mu}(\theta) = \frac{1}{N}\sum_{i=1}^{N} S_i, \quad \hat{\Sigma}(\theta) = \frac{1}{N-1}\sum_{i=1}^{N}(S_i - \hat{\mu}(\theta))(S_i - \hat{\mu}(\theta))^T$$

and use as the likelihood $\phi(S_{obs}; \hat{\mu}(\theta), \hat{\Sigma}(\theta))$ where $\phi(z; \mu, \Sigma)$ is the multivariate normal density with mean μ and covariance Σ.

Synthetic likelihood

Wood, 2010

- There is a curse of dimensionality that is inherent in ABC methods.
  - The ABC likelihood is in effect based on a kernel estimation of the summary statistic distribution.
  - When the number of summary statistics is large, this becomes impractical.
- An alternative likelihood free methodology is synthetic likelihood.

Synthetic likelihood

At each θ, simulate $S_1, \ldots, S_N \sim p(S|\theta)$,

$$\hat{\mu}(\theta) = \frac{1}{N}\sum_{i=1}^{N} S_i, \quad \hat{\Sigma}(\theta) = \frac{1}{N-1}\sum_{i=1}^{N}(S_i - \hat{\mu}(\theta))(S_i - \hat{\mu}(\theta))^T$$

and use as the likelihood $\phi(S_{obs}; \hat{\mu}(\theta), \hat{\Sigma}(\theta))$ where $\phi(z; \mu, \Sigma)$ is the multivariate normal density with mean μ and covariance Σ.

Synthetic likelihood
Wood, 2010

- Some advantages of synthetic likelihood compared to ABC:
  - It mitigates the curse of dimensionality in ABC to some extent, by making parametric assumptions about the distribution of S in likelihood approximations.
  - It's tuning parameters (the number of model simulations N) are easy to set. The tolerance ϵ in ABC is hard to set.
- On the other hand ... normality of $S|\theta$ is assumed.
  - S can often be chosen so that it satisfies some central limit theorem.
  - The approximation of normality can often be improved by transformation.
  - The synthetic likelihood based posterior is often not sensitive to violations of normality (Price *et al.*, 2017).

Synthetic likelihood

Wood, 2010

- Some advantages of synthetic likelihood compared to ABC:
  - It mitigates the curse of dimensionality in ABC to some extent, by making parametric assumptions about the distribution of S in likelihood approximations.
  - It's tuning parameters (the number of model simulations N) are easy to set. The tolerance ϵ in ABC is hard to set.
- On the other hand ... normality of $S|\theta$ is assumed.
  - S can often be chosen so that it satisfies some central limit theorem.
  - The approximation of normality can often be improved by transformation.
  - The synthetic likelihood based posterior is often not sensitive to violations of normality (Price *et al.*, 2017).

Synthetic likelihood
Wood, 2010

- Some advantages of synthetic likelihood compared to ABC:
  - It mitigates the curse of dimensionality in ABC to some extent, by making parametric assumptions about the distribution of S in likelihood approximations.
  - It's tuning parameters (the number of model simulations N) are easy to set. The tolerance ϵ in ABC is hard to set.
- On the other hand ... normality of $S|\theta$ is assumed.
  - S can often be chosen so that it satisfies some central limit theorem.
  - The approximation of normality can often be improved by transformation.
  - The synthetic likelihood based posterior is often not sensitive to violations of normality (Price *et al.*, 2017).

Synthetic likelihood
Wood, 2010

- Some advantages of synthetic likelihood compared to ABC:
  - It mitigates the curse of dimensionality in ABC to some extent, by making parametric assumptions about the distribution of S in likelihood approximations.
  - It's tuning parameters (the number of model simulations N) are easy to set. The tolerance ϵ in ABC is hard to set.
- On the other hand ... normality of $S|\theta$ is assumed.
  - S can often be chosen so that it satisfies some central limit theorem.
  - The approximation of normality can often be improved by transformation.
  - The synthetic likelihood based posterior is often not sensitive to violations of normality (Price *et al.*, 2017).

Synthetic likelihood

Wood, 2010

- Some advantages of synthetic likelihood compared to ABC:
  - It mitigates the curse of dimensionality in ABC to some extent, by making parametric assumptions about the distribution of S in likelihood approximations.
  - It's tuning parameters (the number of model simulations N) are easy to set. The tolerance ϵ in ABC is hard to set.
- On the other hand ... normality of $S|\theta$ is assumed.
  - S can often be chosen so that it satisfies some central limit theorem.
  - The approximation of normality can often be improved by transformation.
  - The synthetic likelihood based posterior is often not sensitive to violations of normality (Price *et al.*, 2017).

Synthetic likelihood

Wood, 2010

- Some advantages of synthetic likelihood compared to ABC:
  - It mitigates the curse of dimensionality in ABC to some extent, by making parametric assumptions about the distribution of S in likelihood approximations.
  - It's tuning parameters (the number of model simulations N) are easy to set. The tolerance ϵ in ABC is hard to set.
- On the other hand ... normality of $S|\theta$ is assumed.
  - S can often be chosen so that it satisfies some central limit theorem.
  - The approximation of normality can often be improved by transformation.
  - The synthetic likelihood based posterior is often not sensitive to violations of normality (Price *et al.*, 2017).

Synthetic likelihood

Wood, 2010

- Some advantages of synthetic likelihood compared to ABC:
  - It mitigates the curse of dimensionality in ABC to some extent, by making parametric assumptions about the distribution of S in likelihood approximations.
  - It's tuning parameters (the number of model simulations N) are easy to set. The tolerance ϵ in ABC is hard to set.
- On the other hand ... normality of $S|\theta$ is assumed.
  - S can often be chosen so that it satisfies some central limit theorem.
  - The approximation of normality can often be improved by transformation.
  - The synthetic likelihood based posterior is often not sensitive to violations of normality (Price *et al.*, 2017).

Synthetic likelihood

Wood, 2010

- Some advantages of synthetic likelihood compared to ABC:
  - It mitigates the curse of dimensionality in ABC to some extent, by making parametric assumptions about the distribution of S in likelihood approximations.
  - It's tuning parameters (the number of model simulations N) are easy to set. The tolerance ϵ in ABC is hard to set.
- On the other hand ... normality of $S|\theta$ is assumed.
  - S can often be chosen so that it satisfies some central limit theorem.
  - The approximation of normality can often be improved by transformation.
  - The synthetic likelihood based posterior is often not sensitive to violations of normality (Price *et al.*, 2017).

Synthetic likelihood
Wood, 2010

- Some refinements:
  - When using synthetic likelihood within MCMC, the targeted distribution does not seem to depend strongly on N.
  - Unbiased likelihood estimation under normality (Ghurye and Olkin, 1969) allowing pseudo-marginal algorithms (Price *et al.*, 2016).
- Shrinkage estimation of covariance with the graphical lasso to reduce the needed number of simulations (An *et al.*, 2016).
- Bootstrap methods for the covariance estimation in computationally expensive models (Everitt, 2017).
- Variational inference methods less sensitive to noise in likelihood estimates (Ong *et al.*, 2018).

Synthetic likelihood

Wood, 2010

- Some refinements:
  - When using synthetic likelihood within MCMC, the targeted distribution does not seem to depend strongly on N.
  - Unbiased likelihood estimation under normality (Ghurye and Olkin, 1969) allowing pseudo-marginal algorithms (Price *et al.*, 2016).
- Shrinkage estimation of covariance with the graphical lasso to reduce the needed number of simulations (An *et al.*, 2016).
- Bootstrap methods for the covariance estimation in computationally expensive models (Everitt, 2017).
- Variational inference methods less sensitive to noise in likelihood estimates (Ong *et al.*, 2018).

Synthetic likelihood
Wood, 2010

- Some refinements:
  - When using synthetic likelihood within MCMC, the targeted distribution does not seem to depend strongly on N.
  - Unbiased likelihood estimation under normality (Ghurye and Olkin, 1969) allowing pseudo-marginal algorithms (Price *et al.*, 2016).
- Shrinkage estimation of covariance with the graphical lasso to reduce the needed number of simulations (An *et al.*, 2016).
- Bootstrap methods for the covariance estimation in computationally expensive models (Everitt, 2017).
- Variational inference methods less sensitive to noise in likelihood estimates (Ong *et al.*, 2018).

Synthetic likelihood

Wood, 2010

- Some refinements:
  - When using synthetic likelihood within MCMC, the targeted distribution does not seem to depend strongly on N.
  - Unbiased likelihood estimation under normality (Ghurye and Olkin, 1969) allowing pseudo-marginal algorithms (Price *et al.*, 2016).
- Shrinkage estimation of covariance with the graphical lasso to reduce the needed number of simulations (An *et al.*, 2016).
- Bootstrap methods for the covariance estimation in computationally expensive models (Everitt, 2017).
- Variational inference methods less sensitive to noise in likelihood estimates (Ong *et al.*, 2018).

Synthetic likelihood

Wood, 2010

- Some refinements:
  - When using synthetic likelihood within MCMC, the targeted distribution does not seem to depend strongly on N.
  - Unbiased likelihood estimation under normality (Ghurye and Olkin, 1969) allowing pseudo-marginal algorithms (Price *et al.*, 2016).
- Shrinkage estimation of covariance with the graphical lasso to reduce the needed number of simulations (An *et al.*, 2016).
- Bootstrap methods for the covariance estimation in computationally expensive models (Everitt, 2017).
- Variational inference methods less sensitive to noise in likelihood estimates (Ong *et al.*, 2018).

Synthetic likelihood

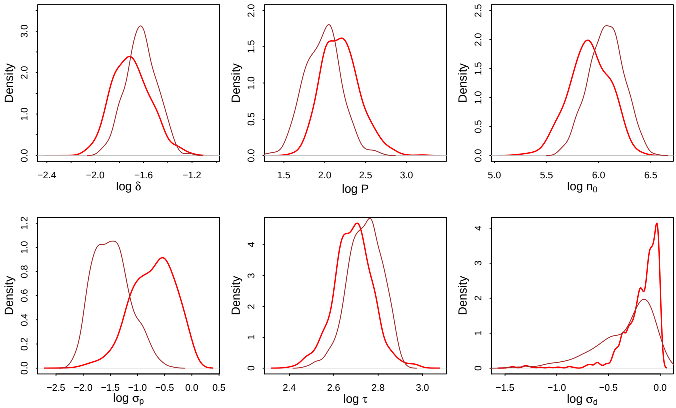
Wood, 2010

- Some refinements:
  - When using synthetic likelihood within MCMC, the targeted distribution does not seem to depend strongly on N.
  - Unbiased likelihood estimation under normality (Ghurye and Olkin, 1969) allowing pseudo-marginal algorithms (Price *et al.*, 2016).
- Shrinkage estimation of covariance with the graphical lasso to reduce the needed number of simulations (An *et al.*, 2016).
- Bootstrap methods for the covariance estimation in computationally expensive models (Everitt, 2017).
- Variational inference methods less sensitive to noise in likelihood estimates (Ong *et al.*, 2018).

Nicholson's Blowflies

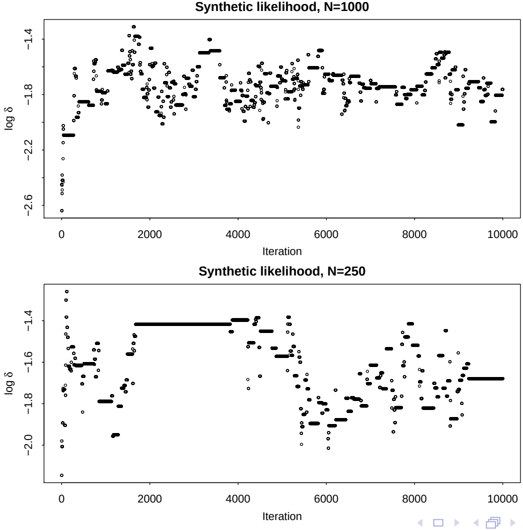
Nicholson (1954, 1957)

Synthetic likelihood estimated marginals (red) with N=1000 and rejection ABC estimated marginals (brown) with NN regression adjustment. Synthetic likelihood uses all 23 summary statistics.



Nicholson's Blowflies
Nicholson (1954,1957)

Trace plots for $\log \delta$ for random walk MCMC with N=1000 and N=250



Synthetic likelihood extensions

- Extensions to address non-normality of summaries - extended empirical saddlepoint approximation (Fasiolo *et al.*, 2016), ratio estimation (Dutta *et al.*, 2016)
- Replace the intractable likelihood with some other pseudo-likelihood:
  - Empirical likelihood (Mengersen, Pudlo and Robert, 2013, Pham 2016).
  - Bootstrap likelihood (Zhu *et al.*, 2016).
  - More general parametric auxiliary models (Drovandi *et al.*, 2015).

Synthetic likelihood extensions

- Extensions to address non-normality of summaries - extended empirical saddlepoint approximation (Fasiolo *et al.*, 2016), ratio estimation (Dutta *et al.*, 2016)
- Replace the intractable likelihood with some other pseudo-likelihood:
  - Empirical likelihood (Mengersen, Pudlo and Robert, 2013, Pham 2016).
  - Bootstrap likelihood (Zhu *et al.*, 2016).
  - More general parametric auxiliary models (Drovandi *et al.*, 2015).

Synthetic likelihood extensions

- Extensions to address non-normality of summaries - extended empirical saddlepoint approximation (Fasiolo *et al.*, 2016), ratio estimation (Dutta *et al.*, 2016)
- Replace the intractable likelihood with some other pseudo-likelihood:
  - Empirical likelihood (Mengersen, Pudlo and Robert, 2013, Pham 2016).
  - Bootstrap likelihood (Zhu *et al.*, 2016).
  - More general parametric auxiliary models (Drovandi *et al.*, 2015).

Synthetic likelihood extensions

- Extensions to address non-normality of summaries - extended empirical saddlepoint approximation (Fasiolo *et al.*, 2016), ratio estimation (Dutta *et al.*, 2016)
- Replace the intractable likelihood with some other pseudo-likelihood:
  - Empirical likelihood (Mengersen, Pudlo and Robert, 2013, Pham 2016).
  - Bootstrap likelihood (Zhu *et al.*, 2016).
  - More general parametric auxiliary models (Drovandi *et al.*, 2015).

Synthetic likelihood extensions

- Extensions to address non-normality of summaries - extended empirical saddlepoint approximation (Fasiolo *et al.*, 2016), ratio estimation (Dutta *et al.*, 2016)
- Replace the intractable likelihood with some other pseudo-likelihood:
  - Empirical likelihood (Mengersen, Pudlo and Robert, 2013, Pham 2016).
  - Bootstrap likelihood (Zhu *et al.*, 2016).
  - More general parametric auxiliary models (Drovandi *et al.*, 2015).

Variational Approximation

Ormerod and Wand, 2010, Blei *et al.*, 2017

Let $q_\lambda(\theta)$ be an approximation to $p(\theta|y_{\text{obs}})$, parametrized by variational parameters λ to be chosen to give the "best" approximation.

The Kullback-Leibler (KL) divergence from $q_\lambda(\theta)$ to $p(\theta|y_{\text{obs}})$, $KL(q_\lambda(\theta)||p(\theta|y_{\text{obs}}))$ can be expressed as

$$\int \log \frac{q_\lambda(\theta)}{p(\theta|y_{\text{obs}})} q_\lambda(\theta)\, d\theta = \log p(y_{\text{obs}}) - \int \log \frac{p(\theta)p(y_{\text{obs}}|\theta)}{q_\lambda(\theta)} q_\lambda(\theta)\, d\theta$$

The second term on the right hand side is the variational lower bound which we write $\mathcal{L}(\lambda)$.

Minimizing $KL(q_\lambda(\theta)||p(\theta|y)$ with respect to λ is the same as maximizing $\mathcal{L}(\lambda)$.

Variational Approximation

Ormerod and Wand, 2010, Blei *et al.*, 2017

Let $q_\lambda(\theta)$ be an approximation to $p(\theta|y_{\text{obs}})$, parametrized by variational parameters λ to be chosen to give the "best" approximation.

The Kullback-Leibler (KL) divergence from $q_\lambda(\theta)$ to $p(\theta|y_{\text{obs}})$, $KL(q_\lambda(\theta)||p(\theta|y_{\text{obs}}))$ can be expressed as

$$\int \log \frac{q_\lambda(\theta)}{p(\theta|y_{\text{obs}})} q_\lambda(\theta)\, d\theta = \log p(y_{\text{obs}}) - \int \log \frac{p(\theta)p(y_{\text{obs}}|\theta)}{q_\lambda(\theta)} q_\lambda(\theta)\, d\theta$$

The second term on the right hand side is the variational lower bound which we write $\mathcal{L}(\lambda)$.

Minimizing $KL(q_\lambda(\theta)||p(\theta|y)$ with respect to λ is the same as maximizing $\mathcal{L}(\lambda)$.

Variational Approximation

Ormerod and Wand, 2010, Blei *et al.*, 2017

Let $q_\lambda(\theta)$ be an approximation to $p(\theta|y_{\text{obs}})$, parametrized by variational parameters λ to be chosen to give the "best" approximation.

The Kullback-Leibler (KL) divergence from $q_\lambda(\theta)$ to $p(\theta|y_{\text{obs}})$, $KL(q_\lambda(\theta)||p(\theta|y_{\text{obs}}))$ can be expressed as

$$\int \log \frac{q_\lambda(\theta)}{p(\theta|y_{\text{obs}})} q_\lambda(\theta)\, d\theta = \log p(y_{\text{obs}}) - \int \log \frac{p(\theta)p(y_{\text{obs}}|\theta)}{q_\lambda(\theta)} q_\lambda(\theta)\, d\theta$$

The second term on the right hand side is the variational lower bound which we write $\mathcal{L}(\lambda)$.

Minimizing $KL(q_\lambda(\theta)||p(\theta|y)$ with respect to λ is the same as maximizing $\mathcal{L}(\lambda)$.

Variational Approximation

Ormerod and Wand, 2010, Blei *et al.*, 2017

Let $q_\lambda(\theta)$ be an approximation to $p(\theta|y_{\text{obs}})$, parametrized by variational parameters λ to be chosen to give the "best" approximation.

The Kullback-Leibler (KL) divergence from $q_\lambda(\theta)$ to $p(\theta|y_{\text{obs}})$, $KL(q_\lambda(\theta)||p(\theta|y_{\text{obs}}))$ can be expressed as

$$\int \log \frac{q_\lambda(\theta)}{p(\theta|y_{\text{obs}})} q_\lambda(\theta)\, d\theta = \log p(y_{\text{obs}}) - \int \log \frac{p(\theta)p(y_{\text{obs}}|\theta)}{q_\lambda(\theta)} q_\lambda(\theta)\, d\theta$$

The second term on the right hand side is the variational lower bound which we write $\mathcal{L}(\lambda)$.

Minimizing $KL(q_\lambda(\theta)||p(\theta|y)$ with respect to λ is the same as maximizing $\mathcal{L}(\lambda)$.

Variational Approximation and ABC

Can variational approximation be used for ABC?

Tran, Nott and Kohn (2017) use an idea similar to pseudo-marginal Metropolis-Hastings.

Suppose we can estimate the likelihood unbiasedly (for ABC we can) by a non-negative estimator $\hat{p}(y_{\text{obs}}|\theta)$.

Write $z = \log \hat{p}(y_{\text{obs}}|\theta) - \log p(y_{\text{obs}}|\theta)$ so that

$$e^z = \frac{\hat{p}(y_{\text{obs}}|\theta)}{p(y_{\text{obs}}|\theta)}.$$

Variational Approximation and ABC

Can variational approximation be used for ABC?

Tran, Nott and Kohn (2017) use an idea similar to pseudo-marginal Metropolis-Hastings.

Suppose we can estimate the likelihood unbiasedly (for ABC we can) by a non-negative estimator $\hat{p}(y_{\text{obs}}|\theta)$.

Write $z = \log \hat{p}(y_{\text{obs}}|\theta) - \log p(y_{\text{obs}}|\theta)$ so that

$$e^z = \frac{\hat{p}(y_{\text{obs}}|\theta)}{p(y_{\text{obs}}|\theta)}.$$

Variational Approximation and ABC

Can variational approximation be used for ABC?

Tran, Nott and Kohn (2017) use an idea similar to pseudo-marginal Metropolis-Hastings.

Suppose we can estimate the likelihood unbiasedly (for ABC we can) by a non-negative estimator $\hat{p}(y_{\text{obs}}|\theta)$.

Write $z = \log \hat{p}(y_{\text{obs}}|\theta) - \log p(y_{\text{obs}}|\theta)$ so that

$$e^z = \frac{\hat{p}(y_{\text{obs}}|\theta)}{p(y_{\text{obs}}|\theta)}.$$

Variational Approximation and ABC

Can variational approximation be used for ABC?

Tran, Nott and Kohn (2017) use an idea similar to pseudo-marginal Metropolis-Hastings.

Suppose we can estimate the likelihood unbiasedly (for ABC we can) by a non-negative estimator $\hat{p}(y_{\text{obs}}|\theta)$.

Write $z = \log \hat{p}(y_{\text{obs}}|\theta) - \log p(y_{\text{obs}}|\theta)$ so that

$$e^z = \frac{\hat{p}(y_{\text{obs}}|\theta)}{p(y_{\text{obs}}|\theta)}.$$

Variational Bayes with intractable likelihood (VBIL)

Tran, Nott and Kohn, 2017

Write $\pi(\theta)$ for the posterior distribution. Consider a joint distribution on (θ, z) of

$$\pi(\theta, z) = \pi(\theta)\exp(z)g(z|\theta),$$

where $g(z|\theta)$ is defined implicitly by the process of generating $\hat{p}(y_{\text{obs}}|\theta)$.

The θ marginal of $\pi(\theta, z)$ is $\pi(\theta)$ since

$$\int \exp(z)g(z|\theta)dz = \int \frac{\hat{p}(y|\theta)}{p(y|\theta)} g(z|\theta)dz = 1$$

by the unbiasedness of $\hat{p}(y|\theta)$.

Variational Bayes with intractable likelihood (VBIL)

Tran, Nott and Kohn, 2017

Write $\pi(\theta)$ for the posterior distribution. Consider a joint distribution on (θ, z) of

$$\pi(\theta, z) = \pi(\theta)\exp(z)g(z|\theta),$$

where $g(z|\theta)$ is defined implicitly by the process of generating $\hat{p}(y_{\text{obs}}|\theta)$.

The θ marginal of $\pi(\theta, z)$ is $\pi(\theta)$ since

$$\int \exp(z)g(z|\theta)dz = \int \frac{\hat{p}(y|\theta)}{p(y|\theta)}g(z|\theta)dz = 1$$

by the unbiasedness of $\hat{p}(y|\theta)$.

Variational Bayes with intractable likelihood (VBIL)

Tran, Nott and Kohn, 2017

Consider

$$q_\lambda(\theta, z) = q_\lambda(\theta) g(z|\theta),$$

for some parametrized approximation $q_\lambda(\theta)$ (for example multivariate normal) to $\pi(\theta)$.

The θ marginal of $q_\lambda(\theta, z)$ is $q_\lambda(\theta)$.

If we can tune the estimator $\hat{p}(y|\theta)$ such that $E(z|\theta)$ does not depend on θ then the variational optimization matching $q_\lambda(\theta, z)$ to $\pi(\theta, z)$ is equivalent to the one matching $q_\lambda(\theta)$ to $\pi(\theta)$.

Variational Bayes with intractable likelihood (VBIL)

Tran, Nott and Kohn, 2017

Consider

$$q_\lambda(\theta, z) = q_\lambda(\theta)g(z|\theta),$$

for some parametrized approximation $q_\lambda(\theta)$ (for example multivariate normal) to $\pi(\theta)$.

The θ marginal of $q_\lambda(\theta, z)$ is $q_\lambda(\theta)$.

If we can tune the estimator $\hat{p}(y|\theta)$ such that $E(z|\theta)$ does not depend on θ then the variational optimization matching $q_\lambda(\theta, z)$ to $\pi(\theta, z)$ is equivalent to the one matching $q_\lambda(\theta)$ to $\pi(\theta)$.

Variational Bayes with intractable likelihood (VBIL)

Tran, Nott and Kohn, 2017

Consider

$$q_\lambda(\theta, z) = q_\lambda(\theta)g(z|\theta),$$

for some parametrized approximation $q_\lambda(\theta)$ (for example multivariate normal) to $\pi(\theta)$.

The θ marginal of $q_\lambda(\theta, z)$ is $q_\lambda(\theta)$.

If we can tune the estimator $\hat{p}(y|\theta)$ such that $E(z|\theta)$ does not depend on θ then the variational optimization matching $q_\lambda(\theta, z)$ to $\pi(\theta, z)$ is equivalent to the one matching $q_\lambda(\theta)$ to $\pi(\theta)$.

Variational Bayes with intractable likelihood (VBIL)
Tran, Nott and Kohn, 2017

Tran, Nott and Kohn (2017) show that the gradient of the lower bound for optimization for $q_\lambda(\theta, z)$ can be estimated unbiasedly and use stochastic gradient methods for the optimization.

VBIL can be used not just in ABC but other applications (eg. state space models). The main advantage over alternatives is less sensitivity to noise in likelihood estimates.

Refinements:

- Reduced variance gradient estimation for models coded in an automatic differentiation environment with model simulations a smooth function of underlying pseudo-random numbers (Moreno *et al.*, 2016).
- Variational methods with the synthetic likelihood (Ong *et al.*, 2017).

Variational Bayes with intractable likelihood (VBIL)

Tran, Nott and Kohn, 2017

Tran, Nott and Kohn (2017) show that the gradient of the lower bound for optimization for $q_\lambda(\theta, z)$ can be estimated unbiasedly and use stochastic gradient methods for the optimization.

VBIL can be used not just in ABC but other applications (eg. state space models). The main advantage over alternatives is less sensitivity to noise in likelihood estimates.

Refinements:

- Reduced variance gradient estimation for models coded in an automatic differentiation environment with model simulations a smooth function of underlying pseudo-random numbers (Moreno *et al.*, 2016).
- Variational methods with the synthetic likelihood (Ong *et al.*, 2017).

Variational Bayes with intractable likelihood (VBIL)

Tran, Nott and Kohn, 2017

Tran, Nott and Kohn (2017) show that the gradient of the lower bound for optimization for $q_\lambda(\theta, z)$ can be estimated unbiasedly and use stochastic gradient methods for the optimization.

VBIL can be used not just in ABC but other applications (eg. state space models). The main advantage over alternatives is less sensitivity to noise in likelihood estimates.

Refinements:

- Reduced variance gradient estimation for models coded in an automatic differentiation environment with model simulations a smooth function of underlying pseudo-random numbers (Moreno *et al.*, 2016).
- Variational methods with the synthetic likelihood (Ong *et al.*, 2017).

Variational Bayes with intractable likelihood (VBIL)

Tran, Nott and Kohn, 2017

Tran, Nott and Kohn (2017) show that the gradient of the lower bound for optimization for $q_\lambda(\theta, z)$ can be estimated unbiasedly and use stochastic gradient methods for the optimization.

VBIL can be used not just in ABC but other applications (eg. state space models). The main advantage over alternatives is less sensitivity to noise in likelihood estimates.

Refinements:

- Reduced variance gradient estimation for models coded in an automatic differentiation environment with model simulations a smooth function of underlying pseudo-random numbers (Moreno *et al.*, 2016).
- Variational methods with the synthetic likelihood (Ong *et al.*, 2017).

Expectation propagation
Minka, 2001

Suppose $h(\theta)$ is a density to be approximated in the form

$$h(\theta) \propto \prod_{i=0}^{n} h_i(\theta).$$

For example

$$\begin{aligned} p(\theta|y_{\text{obs}}) &\propto p(\theta)p(y_{\text{obs}}|\theta) \\ &= p(\theta)\prod_{i=1}^{n} p(y_{\text{obs},i}|y_{\text{obs},<i},\theta). \end{aligned}$$

Expectation propagation

Minka, 2001

Consider an approximation of the same form as the target

$$q(\theta) = \prod_{i=0}^{n} q_i(\theta),$$

where we want $q_i(\theta) \propto h_i(\theta)$.

Replace the current $q_i(\theta)$ by the corresponding term $h_i(\theta)$ in the target to create the tilted distribution

$$\tilde{q}(\theta) = h_i(\theta) \prod_{j \neq i} q_j(\theta),$$

an unnormalized approximation to $h(\theta)$ that should be closer to it than the current $q(\theta)$.

Expectation propagation
Minka, 2001

Consider an approximation of the same form as the target

$$q(\theta) = \prod_{i=0}^{n} q_i(\theta),$$

where we want $q_i(\theta) \propto h_i(\theta)$.

Replace the current $q_i(\theta)$ by the corresponding term $h_i(\theta)$ in the target to create the tilted distribution

$$\tilde{q}(\theta) = h_i(\theta) \prod_{j \neq i} q_j(\theta),$$

an unnormalized approximation to $h(\theta)$ that should be closer to it than the current $q(\theta)$.

Expectation-propagation ABC

Barthelmé and Chopin, 2014

Optimize $q_i(\theta)$ in $q(\theta)$ to get closer to $\tilde{q}(\theta)$ in the Kullback-Leibler sense, cycle over i until convergence (hopefully).

Barthelmé and Chopin (2014) apply EP to ABC: consider the

EP-ABC likelihood approximation

$$p(y_{\text{obs}}|\theta) = \prod_{i=1}^{n} \int p(y_i|y_{\text{obs},<i},\theta) I(d(y_i, y_{\text{obs},i}) < \epsilon) dy_i.$$

This likelihood approximation has the form of a product and we can approximate the terms of it using EP.

Expectation-propagation ABC

Barthelmé and Chopin, 2014

Optimize $q_i(\theta)$ in $q(\theta)$ to get closer to $\tilde{q}(\theta)$ in the Kullback-Leibler sense, cycle over i until convergence (hopefully).

Barthelmé and Chopin (2014) apply EP to ABC: consider the

EP-ABC likelihood approximation

$$p(y_{\text{obs}}|\theta) = \prod_{i=1}^{n} \int p(y_i|y_{\text{obs},<i}, \theta) I(d(y_i, y_{\text{obs},i}) < \epsilon) dy_i.$$

This likelihood approximation has the form of a product and we can approximate the terms of it using EP.

Expectation-propagation ABC

Barthelmé and Chopin, 2014

The big advantage of this technique is that no summary statistics are required.

Possible drawbacks:

- To do the moment matching steps of EP-ABC it must be possible to simulate from $p(y_i|y_{\text{obs},<i},\theta)$.
- We need an exponential family form (usually normal) for the approximation.
- EP is generally not guaranteed to converge.
- The EP optimization steps are done stochastically in EP-ABC and sophisticated variance reduction methods may be needed to ensure stability.

Expectation-propagation ABC

Barthelmé and Chopin, 2014

The big advantage of this technique is that no summary statistics are required.

Possible drawbacks:

- To do the moment matching steps of EP-ABC it must be possible to simulate from $p(y_i|y_{\text{obs},<i},\theta)$.
- We need an exponential family form (usually normal) for the approximation.
- EP is generally not guaranteed to converge.
- The EP optimization steps are done stochastically in EP-ABC and sophisticated variance reduction methods may be needed to ensure stability.

Expectation-propagation ABC

Barthelmé and Chopin, 2014

The big advantage of this technique is that no summary statistics are required.

Possible drawbacks:

- To do the moment matching steps of EP-ABC it must be possible to simulate from $p(y_i|y_{\text{obs},<i},\theta)$.
- We need an exponential family form (usually normal) for the approximation.
- EP is generally not guaranteed to converge.
- The EP optimization steps are done stochastically in EP-ABC and sophisticated variance reduction methods may be needed to ensure stability.

Expectation-propagation ABC

Barthelmé and Chopin, 2014

The big advantage of this technique is that no summary statistics are required.

Possible drawbacks:

- To do the moment matching steps of EP-ABC it must be possible to simulate from $p(y_i|y_{\text{obs},<i}, \theta)$.
- We need an exponential family form (usually normal) for the approximation.
- EP is generally not guaranteed to converge.
- The EP optimization steps are done stochastically in EP-ABC and sophisticated variance reduction methods may be needed to ensure stability.

Expectation-propagation ABC

Barthelmé and Chopin, 2014

The big advantage of this technique is that no summary statistics are required.

Possible drawbacks:

- To do the moment matching steps of EP-ABC it must be possible to simulate from $p(y_i|y_{\text{obs},<i},\theta)$.
- We need an exponential family form (usually normal) for the approximation.
- EP is generally not guaranteed to converge.
- The EP optimization steps are done stochastically in EP-ABC and sophisticated variance reduction methods may be needed to ensure stability.

Neglected topics

- Theory (Frazier *et al.*, 2018).
- ABC model choice (Marin *et al.*, 2018).
- Methods for computationally expensive simulation models (Gutmann and Corander, 2016, Prangle, 2016, Holden *et al.*, 2018).
- Regression based approaches based on mixtures, random forests, deep learning methods, kernels (Bonassi *et al.*, 2011, Fan, Nott and Sisson, 2013, Marin *et al.*, 2016, Jiang *et al.*, 2017, Nakagone *et al.*, 2013).
- Recent high-dimensional ABC methods based on marginal adjustments and copulas (Nott *et al.*, 2014, Li *et al.*, 2017).
- Software - for the examples in this talk I've used various R packages - `abc` (Csillery *et al.*, 2012) `abctools` (Nunes and Prangle, 2015), `EasyABC` (Jabot *et al.*, 2015), `synlik` (Fasiolo and Wood, 2014).

Neglected topics

- Theory (Frazier *et al.*, 2018).
- ABC model choice (Marin *et al.*, 2018).
- Methods for computationally expensive simulation models (Gutmann and Corander, 2016, Prangle, 2016, Holden *et al.*, 2018).
- Regression based approaches based on mixtures, random forests, deep learning methods, kernels (Bonassi *et al.*, 2011, Fan, Nott and Sisson, 2013, Marin *et al.*, 2016, Jiang *et al.*, 2017, Nakagone *et al.*, 2013).
- Recent high-dimensional ABC methods based on marginal adjustments and copulas (Nott *et al.*, 2014, Li *et al.*, 2017).
- Software - for the examples in this talk I've used various R packages - `abc` (Csillery *et al.*, 2012) `abctools` (Nunes and Prangle, 2015), `EasyABC` (Jabot *et al.*, 2015), `synlik` (Fasiolo and Wood, 2014).

Neglected topics

- Theory (Frazier *et al.*, 2018).
- ABC model choice (Marin *et al.*, 2018).
- Methods for computationally expensive simulation models (Gutmann and Corander, 2016, Prangle, 2016, Holden *et al.*, 2018).
- Regression based approaches based on mixtures, random forests, deep learning methods, kernels (Bonassi *et al.*, 2011, Fan, Nott and Sisson, 2013, Marin *et al.*, 2016, Jiang *et al.*, 2017, Nakagone *et al.*, 2013).
- Recent high-dimensional ABC methods based on marginal adjustments and copulas (Nott *et al.*, 2014, Li *et al.*, 2017).
- Software - for the examples in this talk I've used various R packages - `abc` (Csillery *et al.*, 2012) `abctools` (Nunes and Prangle, 2015), `EasyABC` (Jabot *et al.*, 2015), `synlik` (Fasiolo and Wood, 2014).

Neglected topics

- Theory (Frazier *et al.*, 2018).
- ABC model choice (Marin *et al.*, 2018).
- Methods for computationally expensive simulation models (Gutmann and Corander, 2016, Prangle, 2016, Holden *et al.*, 2018).
- Regression based approaches based on mixtures, random forests, deep learning methods, kernels (Bonassi *et al.*, 2011, Fan, Nott and Sisson, 2013, Marin *et al.*, 2016, Jiang *et al.*, 2017, Nakagone *et al.*, 2013).
- Recent high-dimensional ABC methods based on marginal adjustments and copulas (Nott *et al.*, 2014, Li *et al.*, 2017).
- Software - for the examples in this talk I've used various R packages - `abc` (Csillery *et al.*, 2012) `abctools` (Nunes and Prangle, 2015), `EasyABC` (Jabot *et al.*, 2015), `synlik` (Fasiolo and Wood, 2014).

Neglected topics

- Theory (Frazier *et al.*, 2018).
- ABC model choice (Marin *et al.*, 2018).
- Methods for computationally expensive simulation models (Gutmann and Corander, 2016, Prangle, 2016, Holden *et al.*, 2018).
- Regression based approaches based on mixtures, random forests, deep learning methods, kernels (Bonassi *et al.*, 2011, Fan, Nott and Sisson, 2013, Marin *et al.*, 2016, Jiang *et al.*, 2017, Nakagone *et al.*, 2013).
- Recent high-dimensional ABC methods based on marginal adjustments and copulas (Nott *et al.*, 2014, Li *et al.*, 2017).
- Software - for the examples in this talk I've used various R packages - `abc` (Csillery *et al.*, 2012) `abctools` (Nunes and Prangle, 2015), `EasyABC` (Jabot *et al.*, 2015), `synlik` (Fasiolo and Wood, 2014).

Neglected topics

- Theory (Frazier *et al.*, 2018).
- ABC model choice (Marin *et al.*, 2018).
- Methods for computationally expensive simulation models (Gutmann and Corander, 2016, Prangle, 2016, Holden *et al.*, 2018).
- Regression based approaches based on mixtures, random forests, deep learning methods, kernels (Bonassi *et al.*, 2011, Fan, Nott and Sisson, 2013, Marin *et al.*, 2016, Jiang *et al.*, 2017, Nakagone *et al.*, 2013).
- Recent high-dimensional ABC methods based on marginal adjustments and copulas (Nott *et al.*, 2014, Li *et al.*, 2017).
- Software - for the examples in this talk I've used various R packages - `abc` (Csillery *et al.*, 2012) `abctools` (Nunes and Prangle, 2015), `EasyABC` (Jabot *et al.*, 2015), `synlik` (Fasiolo and Wood, 2014).

Thank you