Data-driven Model Selection for Approximate Bayesian Computation via Multiple Logisitic Regression.

Ben Rohrlach

Prof. Nigel Bean, Dr Jonathan Tuke

University of Adelaide

November 6, 2014

Table of Contents

THE UNIVERSITY of ADELAIDE

Some motivation.

- Consider the Beringian Steppe Bison.

Some motivation.

Some motivation.

Some motivation.

- Population numbers dropped at *some time* in the past.

Some motivation.

- Population numbers dropped at *some time* in the past.
- Did it happen slowly over time?

Some motivation.

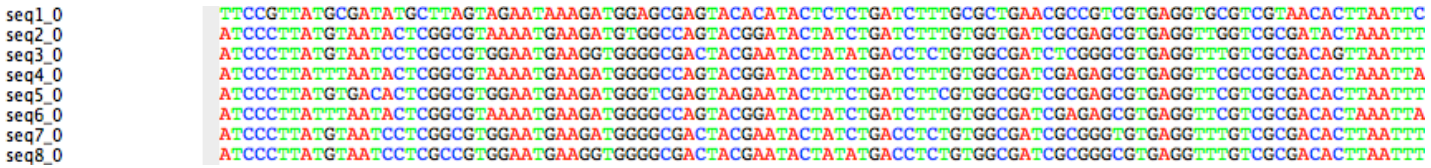
- Population numbers dropped at *some time* in the past.
- Did it happen slowly over time?
- Did it happen abruptly?

Some motivation.

- Population numbers dropped at *some time* in the past.
- Did it happen slowly over time?
- Did it happen abruptly?
- If it did happen abruptly, when did it happen?

Some motivation.

- Population numbers dropped at *some time* in the past.
- Did it happen slowly over time?
- Did it happen abruptly?
- If it did happen abruptly, when did it happen?
- How can we work this out if all we have are some DNA from old bones??



Some motivation.

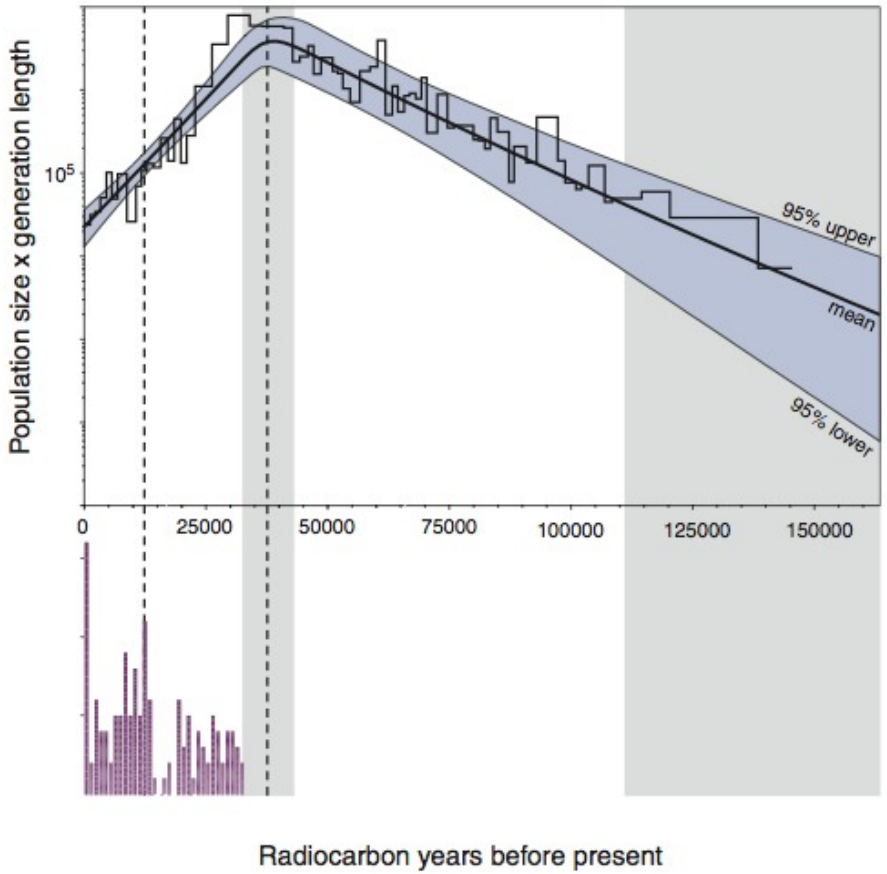


Figure: *Rise and fall of the Beringian steppe bison*, Shapiro *et al.* [4].

Bayesian vs Frequentist.

Frequentist Approach

Bayesian Approach

Bayesian vs Frequentist.

Frequentist Approach

- Data comes from a repeatable experiment.

Bayesian Approach

- Data comes from a realised experiment.

Bayesian vs Frequentist.

Frequentist Approach

- Data comes from a repeatable experiment.
- The parameters are constant.

Bayesian Approach

- Data comes from a realised experiment.
- The parameters are unknown.

Bayesian vs Frequentist.

Frequentist Approach

- Data comes from a repeatable experiment.
- The parameters are constant.
- The parameters are fixed.

Bayesian Approach

- Data comes from a realised experiment.
- The parameters are unknown.
- The data is fixed.

Bayesian vs Frequentist.

In a frequentist analysis we:

- Set α in advance and find $L(\boldsymbol{X} \big| H_0)$,

Bayesian vs Frequentist.

In a frequentist analysis we:

- Set α in advance and find $L(\boldsymbol{X}|H_0)$,
- Accept H_0 if $L(\boldsymbol{X}|H_0) \geq \alpha$,

Bayesian vs Frequentist.

In a frequentist analysis we:

- Set α in advance and find $L(\boldsymbol{X}|H_0)$,
- Accept H_0 if $L(\boldsymbol{X}|H_0) \geq \alpha$,
- Report point estimates and confidence intervals for parameters.

Bayesian vs Frequentist.

In a Bayesian analysis we:

- From $\pi(\boldsymbol{\theta})$ we (inductively) find $P(\boldsymbol{\theta} \big| \boldsymbol{X})$,

Bayesian vs Frequentist.

In a Bayesian analysis we:

- From $\pi(\boldsymbol{\theta})$ we (inductively) find $P(\boldsymbol{\theta}\big|\boldsymbol{X})$,
- Describe the *posterior* distribution of $\boldsymbol{\theta}$,

Bayesian vs Frequentist.

In a Bayesian analysis we:

- From $\pi(\boldsymbol{\theta})$ we (inductively) find $P(\boldsymbol{\theta}\big|\boldsymbol{X})$,
- Describe the *posterior* distribution of $\boldsymbol{\theta}$,
- Report highest posterior density intervals for parameters.

That is:

- We aim to describe the probability of model parameters *given* the data we have observed via

$$P(\theta|\boldsymbol{X}) = \frac{L(\boldsymbol{X}|\theta)\pi(\theta)}{P(\boldsymbol{X})}$$

where $L(\boldsymbol{X}|\theta)$ is the likelihood function for the data.

That is:

- We aim to describe the probability of model parameters *given* the data we have observed via

$$P(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{L(\boldsymbol{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{P(\boldsymbol{X})}$$

where $\pi(\boldsymbol{\theta})$ is the 'prior distribution' for $\boldsymbol{\theta}$ (my prior beliefs about the possible parameter values).

That is:

- We aim to describe the probability of model parameters *given* the data we have observed via

$$P(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{L(\boldsymbol{X}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{P(\boldsymbol{X})}$$

where $P(\boldsymbol{X})$ is the 'marginal likelihood' of the data (sometimes called the 'model evidence').

ABCs History.

- First considered by Donald Rubin in the 1980's via the 'Acceptance-Rejection Algorithm' [1].

ABCs History.

- First considered by Donald Rubin in the 1980's via the 'Acceptance-Rejection Algorithm' [1].
- Particularly useful when obtaining the likelihood function $L(\boldsymbol{X}|\theta)$ is difficult or impossible to obtain.

ABCs History.

- First considered by Donald Rubin in the 1980's via the 'Acceptance-Rejection Algorithm' [1].
- Particularly useful when obtaining the likelihood function $L(\boldsymbol{X}|\theta)$ is difficult or impossible to obtain.
- Relies on being able to simulate data efficiently.

The Rejection-Acceptance Algorithm.

- Consider obtaining ℓ posterior samples using some observed data $\boldsymbol{X}_{obs}$:

1: Set $i = 0$
2: **while** $i < \ell$ **do**

9: **end while**

The Rejection-Acceptance Algorithm.

- Consider obtaining ℓ posterior samples using some observed data $\boldsymbol{X}_{obs}$:

1: Set $i = 0$
2: **while** $i < \ell$ **do**
3: Sample θ^* from $\pi(\theta)$

9: **end while**

The Rejection-Acceptance Algorithm.

- Consider obtaining ℓ posterior samples using some observed data $\boldsymbol{X}_{obs}$:

1: Set $i = 0$
2: **while** $i < \ell$ **do**
3: Sample $\boldsymbol{\theta}^*$ from $\pi(\boldsymbol{\theta})$
4: Simulate $\boldsymbol{X}^*$ from $f(\boldsymbol{X}|\boldsymbol{\theta}^*)$

9: **end while**

The Rejection-Acceptance Algorithm.

- Consider obtaining ℓ posterior samples using some observed data $\boldsymbol{X}_{obs}$:

1: Set $i = 0$
2: **while** $i < \ell$ **do**
3: Sample $\boldsymbol{\theta}^*$ from $\pi(\boldsymbol{\theta})$
4: Simulate $\boldsymbol{X}^*$ from $f(\boldsymbol{X}|\boldsymbol{\theta}^*)$
5: **if** $(\boldsymbol{X}^* = \boldsymbol{X}_{obs})$ **then**

8: **end if**
9: **end while**

The Rejection-Acceptance Algorithm.

- Consider obtaining ℓ posterior samples using some observed data $\boldsymbol{X}_{obs}$:

1: Set $i = 0$
2: **while** $i < \ell$ **do**
3: Sample θ^* from $\pi(\theta)$
4: Simulate $\boldsymbol{X}^*$ from $f(\boldsymbol{X}|\theta^*)$
5: **if** $(\boldsymbol{X}^* = \boldsymbol{X}_{obs})$ **then**
6: accept θ^*
7: $i = i + 1$
8: **end if**
9: **end while**

The Rejection-Acceptance Algorithm.

- Gives the true posterior distribution $P(\theta \big| \boldsymbol{X}_{obs})$.

The Rejection-Acceptance Algorithm.

- Gives the true posterior distribution $P(\theta | \boldsymbol{X}_{obs})$.
- Extremely slow convergence in cases where our data has high dimensionality.

The Rejection-Acceptance Algorithm.

- Gives the true posterior distribution $P(\theta \big| \boldsymbol{X}_{obs})$.
- Extremely slow convergence in cases where our data has high dimensionality.
- Could consider accepting data that is 'close enough'.

The Rejection-Acceptance Algorithm.

- Gives the true posterior distribution $P(\theta | \boldsymbol{X}_{obs})$.
- Extremely slow convergence in cases where our data has high dimensionality.
- Could consider accepting data that is 'close enough'.
- If "$\boldsymbol{X}^* = \boldsymbol{X}_{obs}$" is unrealistic, try "$\boldsymbol{X}^* \approx \boldsymbol{X}_{obs}$"

THE UNIVERSITY of ADELAIDE

The Rejection-Acceptance Algorithm.

- For some distance function $\rho(\boldsymbol{X}, \boldsymbol{Y})$, and some 'tolerance' parameter ϵ, the algorithm now becomes:

The Rejection-Acceptance Algorithm.

- For some distance function $\rho(\boldsymbol{X}, \boldsymbol{Y})$, and some 'tolerance' parameter ϵ, the algorithm now becomes:

1: Set $i = 0$
2: **while** $i < \ell$ **do**
3: Sample $\boldsymbol{\theta}^*$ from $\pi(\boldsymbol{\theta})$
4: Simulate $\boldsymbol{X}^*$ from $f(\boldsymbol{X}|\boldsymbol{\theta})^*$
5: **if** $(\rho(\boldsymbol{X}^*, \boldsymbol{X}_{obs}) < \epsilon)$ **then**
6: accept $\boldsymbol{\theta}^*$
7: $i = i + 1$
8: **end if**
9: **end while**

The Rejection-Acceptance Algorithm.

- Gives an approximate posterior distribution $P(\theta | \hat{\boldsymbol{X}}_{obs})$.

The Rejection-Acceptance Algorithm.

- Gives an approximate posterior distribution $P(\theta | \hat{\boldsymbol{X}}_{obs})$.
- $P(\theta | \hat{\boldsymbol{X}}_{obs}) \to P(\theta | \boldsymbol{X}_{obs})$ as $\epsilon \to 0$.

The Rejection-Acceptance Algorithm.

- Gives an approximate posterior distribution $P(\theta | \hat{\boldsymbol{X}}_{obs})$.
- $P(\theta | \hat{\boldsymbol{X}}_{obs}) \to P(\theta | \boldsymbol{X}_{obs})$ as $\epsilon \to 0$.
- Still slow convergence for small ϵ.

The Rejection-Acceptance Algorithm.

- Gives an approximate posterior distribution $P(\theta | \hat{\boldsymbol{X}}_{obs})$.
- $P(\theta | \hat{\boldsymbol{X}}_{obs}) \to P(\theta | \boldsymbol{X}_{obs})$ as $\epsilon \to 0$.
- Still slow convergence for small ϵ.
- Data being 'similar' can still be very unlikely.

ABC Using Summary Statistics.

What are summary statistics?

ABC Using Summary Statistics.

What are summary statistics?

- A summary statistic is a function of the data (i.e. the sample mean $\bar{\boldsymbol{X}}$).

ABC Using Summary Statistics.

What are summary statistics?

- A summary statistic is a function of the data (i.e. the sample mean $\bar{\boldsymbol{X}}$).
- Summary statistics are used to reduce the dimensionality of data.

ABC Using Summary Statistics.

What are **sufficient** summary statistics?

ABC Using Summary Statistics.

What are **sufficient** summary statistics?

- Sufficient summary statistics contain all of the information about a parameter that is available in a sample (i.e. $\bar{\boldsymbol{X}}$ is sufficient for μ).

ABC Using Summary Statistics.

What are **sufficient** summary statistics?

- Sufficient summary statistics contain all of the information about a parameter that is available in a sample (i.e. $\bar{\boldsymbol{X}}$ is sufficient for μ).
- A summary statistic $S(\boldsymbol{X})$ is sufficient if it can be written in Fisher-Neymann factorised form:

$$L\left(\boldsymbol{X}\middle|\theta\right) = g(\boldsymbol{X})h_\theta\left(S(\boldsymbol{X})\middle|\theta\right)$$

ABC Using Summary Statistics.

- It can be shown $P\left(\boldsymbol{\theta}\middle|\boldsymbol{X}_{obs}\right) = P\left(\boldsymbol{\theta}\middle|S(\boldsymbol{X}_{obs})\right)$.

ABC Using Summary Statistics.

- It can be shown $P\left(\theta \middle| \boldsymbol{X}_{obs}\right) = P\left(\theta \middle| S(\boldsymbol{X}_{obs})\right)$.
- That is, we can compare sufficient summary statistics to obtain the exact posterior distribution for θ.

The Modified Rejection-Acceptance Algorithm.

- For some distance function $\rho(S(\boldsymbol{X}), S(\boldsymbol{Y}))$, and some 'tolerance' parameter ϵ, the algorithm now becomes:

1: Set $i = 0$
2: **while** $i < \ell$ **do**
3: Sample $\boldsymbol{\theta}^*$ from $\pi(\boldsymbol{\theta})$
4: Simulate $\boldsymbol{X}^*$ from $f(\boldsymbol{X}|\boldsymbol{\theta}^*)$
5: **if** $(\rho(S(\boldsymbol{X}^*), S(\boldsymbol{X}_{obs})) < \epsilon)$ **then**
6: accept $\boldsymbol{\theta}^*$
7: $i = i + 1$
8: **end if**
9: **end while**

ABC Using Summary Statistics.

- Gives the same posterior distribution $P(\theta | \hat{S(\boldsymbol{X}_{obs})})$ if $S(\boldsymbol{X})$ is sufficient.

ABC Using Summary Statistics.

- Gives the same posterior distribution $P(\theta | \hat{S(\boldsymbol{X}_{obs})})$ if $S(\boldsymbol{X})$ is sufficient.
- Again, $P(\theta | \hat{S(\boldsymbol{X}_{obs})}) \rightarrow P(\theta | \boldsymbol{X}_{obs})$ as $\epsilon \rightarrow 0$.

ABC Using Summary Statistics.

- Gives the same posterior distribution $P(\theta|\hat{S(\boldsymbol{X}_{obs})})$ if $S(\boldsymbol{X})$ is sufficient.
- Again, $P(\theta|\hat{S(\boldsymbol{X}_{obs})}) \rightarrow P(\theta|\boldsymbol{X}_{obs})$ as $\epsilon \rightarrow 0$.
- Convergence can now be faster.

ABC Using Summary Statistics.

- Gives the same posterior distribution $P(\theta|\hat{S(\boldsymbol{X}_{obs})})$ if $S(\boldsymbol{X})$ is sufficient.
- Again, $P(\theta|\hat{S(\boldsymbol{X}_{obs})}) \rightarrow P(\theta|\boldsymbol{X}_{obs})$ as $\epsilon \rightarrow 0$.
- Convergence can now be faster.
- Sufficient summary statistics rarely show up when required.

ABC Using Summary Statistics.

- Gives the same posterior distribution $P(\theta | \hat{S(\boldsymbol{X}_{obs})})$ if $S(\boldsymbol{X})$ is sufficient.
- Again, $P(\theta | \hat{S(\boldsymbol{X}_{obs})}) \rightarrow P(\theta | \boldsymbol{X}_{obs})$ as $\epsilon \rightarrow 0$.
- Convergence can now be faster.
- Sufficient summary statistics rarely show up when required.
- Choosing a 'best summary statistic' was the focus of my Masters [2].

Approximately Sufficient Summary Statistics

- We have insufficient summary statistics $\boldsymbol{S} = \{S_1, \cdots, S_T\}$.

Approximately Sufficient Summary Statistics

- We have insufficient summary statistics $\boldsymbol{S} = \{S_1, \cdots, S_T\}$.
- We have parameters of interest $\boldsymbol{\Phi} = \{\phi_1, \cdots, \phi_P\}$

Approximately Sufficient Summary Statistics

- We have insufficient summary statistics $\boldsymbol{S} = \{S_1, \cdots, S_T\}$.
- We have parameters of interest $\boldsymbol{\Phi} = \{\phi_1, \cdots, \phi_P\}$
- Create Γ simulations, which gives $\Gamma \times T$ summary statistics with known input parameters (call this TrainDat).

Approximately Sufficient Summary Statistics

- We have insufficient summary statistics $\boldsymbol{S} = \{S_1, \cdots, S_T\}$.
- We have parameters of interest $\boldsymbol{\Phi} = \{\phi_1, \cdots, \phi_P\}$
- Create Γ simulations, which gives $\Gamma \times T$ summary statistics with known input parameters (call this TrainDat).
- For each $n \in \{1, \cdots, P\}$ perform linear regression on the TrainDat such that we can get predictions

$$\hat{\phi}_n = \hat{\beta}_0^{(n)} + \sum_{j=1}^{T} \hat{\beta}_j^{(n)} s_j$$

Approximately Sufficient Summary Statistics

- We have insufficient summary statistics $\boldsymbol{S} = \{S_1, \cdots, S_T\}$.
- We have parameters of interest $\boldsymbol{\Phi} = \{\phi_1, \cdots, \phi_P\}$
- Create Γ simulations, which gives $\Gamma \times T$ summary statistics with known input parameters (call this TrainDat).
- For each $n \in \{1, \cdots, P\}$ perform linear regression on the TrainDat such that we can get predictions

$$\hat{\phi}_n = \hat{\beta}_0^{(n)} + \sum_{j=1}^{T} \hat{\beta}_j^{(n)} s_j$$

- We now have a 'best predicted parameter value' if we have summary statistics.

THE UNIVERSITY of ADELAIDE

Model Selection in ABC.

How do we choose which model we might wish to simulate data under?

Model Selection in ABC.

- Consider models $\mathcal{M} = \{M_1, \cdots, M_q\}$

Model Selection in ABC.

- Consider models $\mathcal{M} = \{M_1, \cdots, M_q\}$
- We can add a step which selects which model we might simulate under.

The Very Modified Rejection-Acceptance Algorithm.

- Let $R(M_k)$ be the probability of Model k, and $\pi_k(\theta)$ be the prior distribution for parameters under Model k.

The Very Modified Rejection-Acceptance Algorithm.

- Let $R(M_k)$ be the probability of Model k, and $\pi_k(\theta)$ be the prior distribution for parameters under Model k.
- Consider obtaining ℓ posterior samples from a possible q models using some observed data $\boldsymbol{X}_{obs}$:

The Very Modified Rejection-Acceptance Algorithm.

- Let $R(M_k)$ be the probability of Model k, and $\pi_k(\theta)$ be the prior distribution for parameters under Model k.
- Consider obtaining ℓ posterior samples from a possible q models using some observed data $\boldsymbol{X}_{obs}$:

1: Set $i = 0$
2: **while** $i < \ell$ **do**

10: **end while**

THE UNIVERSITY of ADELAIDE

The Very Modified Rejection-Acceptance Algorithm.

- Let $R(M_k)$ be the probability of Model k, and $\pi_k(\theta)$ be the prior distribution for parameters under Model k.
- Consider obtaining ℓ posterior samples from a possible q models using some observed data $\boldsymbol{X}_{obs}$:

1: Set $i = 0$
2: **while** $i < \ell$ **do**
3: Randomly select some model k to simulate via $R(\cdot)$

10: **end while**

The Very Modified Rejection-Acceptance Algorithm.

- Let $R(M_k)$ be the probability of Model k, and $\pi_k(\theta)$ be the prior distribution for parameters under Model k.
- Consider obtaining ℓ posterior samples from a possible q models using some observed data $\boldsymbol{X}_{obs}$:

1: Set $i = 0$
2: **while** $i < \ell$ **do**
3: Randomly select some model k to simulate via $R(\cdot)$
4: Sample θ^* from $\pi_k(\theta)$

10: **end while**

The Very Modified Rejection-Acceptance Algorithm.

- Let $R(M_k)$ be the probability of Model k, and $\pi_k(\theta)$ be the prior distribution for parameters under Model k.
- Consider obtaining ℓ posterior samples from a possible q models using some observed data $\boldsymbol{X}_{obs}$:

1: Set $i = 0$
2: **while** $i < \ell$ **do**
3: Randomly select some model k to simulate via $R(\cdot)$
4: Sample $\boldsymbol{\theta}^*$ from $\pi_k(\boldsymbol{\theta})$
5: Simulate $\boldsymbol{X}^*$ from $f_k(\boldsymbol{X}|\boldsymbol{\theta}^*)$
6: **if** $(\rho(S(\boldsymbol{X}^*), S(\boldsymbol{X}_{obs})) < \epsilon)$ **then**
7: accept $\boldsymbol{\theta}^*$
8: $i = i + 1$
9: **end if**
10: **end while**

THE UNIVERSITY of ADELAIDE

Model Selection in ABC.

- How can we choose which M_i best fits our data?

Model Selection in ABC.

- How can we choose which M_i best fits our data?
- Common approach is to use 'Bayes Factors' B_{ij}, $i \neq j \in \{1, \cdots, q\}$.

Bayes Factors.

- The Bayes Factor for Models *i* and *j* is:

Bayes Factors.

- The Bayes Factor for Models i and j is:

$$B_{ij} = \frac{P\left(\boldsymbol{X}\middle|M_i\right)}{P\left(\boldsymbol{X}\middle|M_j\right)}$$

Bayes Factors.

- The Bayes Factor for Models i and j is:

$$B_{ij} = \frac{P\left(\boldsymbol{X}\middle|M_i\right)}{P\left(\boldsymbol{X}\middle|M_j\right)} = \frac{P\left(M_i\middle|\boldsymbol{X}\right)P\left(\boldsymbol{X}\right)/R(M_i)}{P\left(M_j\middle|\boldsymbol{X}\right)P\left(\boldsymbol{X}\right)/R(M_j)}$$

Bayes Factors.

- The Bayes Factor for Models i and j is:

$$\begin{aligned} B_{ij} &= \frac{P\left(\boldsymbol{X}\middle|M_i\right)}{P\left(\boldsymbol{X}\middle|M_j\right)} \\ &= \frac{P\left(M_i\middle|\boldsymbol{X}\right)P\left(\boldsymbol{X}\right)/R(M_i)}{P\left(M_j\middle|\boldsymbol{X}\right)P\left(\boldsymbol{X}\right)/R(M_j)} \\ &= \frac{P\left(M_i\middle|\boldsymbol{X}\right)}{P\left(M_j\middle|\boldsymbol{X}\right)}, \end{aligned}$$

  if $R(\cdot)$ has a uniform distribution.

Bayes Factors.

- The Bayes Factor for Models i and j is

$$B_{ij} = \frac{P\left(M_i \middle| \boldsymbol{X}\right)}{P\left(M_j \middle| \boldsymbol{X}\right)}.$$

Bayes Factors.

- The Bayes Factor for Models i and j is

$$B_{ij} = \frac{P\left(M_i \middle| \boldsymbol{X}\right)}{P\left(M_j \middle| \boldsymbol{X}\right)}.$$

- This is just the 'posterior ratio' for Models i and j.

Bayes Factors.

- The Bayes Factor for Models i and j is

$$B_{ij} = \frac{P\left(M_i \middle| \boldsymbol{X}\right)}{P\left(M_j \middle| \boldsymbol{X}\right)}.$$

- This is just the 'posterior ratio' for Models i and j.
- Imagine out of 300 retained posterior parameter samples: 200 are from model i, and 100 are from model j,
  $\Longrightarrow$ B_{ij}

Bayes Factors.

- The Bayes Factor for Models i and j is

$$B_{ij} = \frac{P\left(M_i \middle| \boldsymbol{X}\right)}{P\left(M_j \middle| \boldsymbol{X}\right)}.$$

- This is just the 'posterior ratio' for Models i and j.
- Imagine out of 300 retained posterior parameter samples: 200 are from model i, and 100 are from model j,
  $\implies B_{ij} = \frac{200/300}{100/300}$

Bayes Factors.

- The Bayes Factor for Models i and j is

$$B_{ij} = \frac{P\left(M_i \middle| \boldsymbol{X}\right)}{P\left(M_j \middle| \boldsymbol{X}\right)}.$$

- This is just the 'posterior ratio' for Models i and j.
- Imagine out of 300 retained posterior parameter samples: 200 are from model i, and 100 are from model j,
  $\implies B_{ij} = \frac{200/300}{100/300} = 2.$

A Fundamental Flaw of Bayes Factors.

- It can be shown that [3]:

$$B_{ij} = \frac{P\left(M_i \middle| \boldsymbol{X}\right)}{P\left(M_j \middle| \boldsymbol{X}\right)} \times \frac{h_j\left(\boldsymbol{X} \middle| S(\boldsymbol{X})\right)}{h_i\left(\boldsymbol{X} \middle| S(\boldsymbol{X})\right)}$$

A Fundamental Flaw of Bayes Factors.

- It can be shown that [3]:

$$B_{ij} = \frac{P\left(M_i \middle| \boldsymbol{X}\right)}{P\left(M_j \middle| \boldsymbol{X}\right)} \times \frac{h_j\left(\boldsymbol{X} \middle| S(\boldsymbol{X})\right)}{h_i\left(\boldsymbol{X} \middle| S(\boldsymbol{X})\right)}$$
$$= \frac{P\left(M_i \middle| \boldsymbol{X}\right)}{P\left(M_j \middle| \boldsymbol{X}\right)}$$

A Fundamental Flaw of Bayes Factors.

- It can be shown that [3]:

$$
\begin{aligned}
B_{ij} &= \frac{P\left(M_i \middle| \boldsymbol{X}\right)}{P\left(M_j \middle| \boldsymbol{X}\right)} \times \frac{h_j\left(\boldsymbol{X} \middle| S(\boldsymbol{X})\right)}{h_i\left(\boldsymbol{X} \middle| S(\boldsymbol{X})\right)} \\
&= \frac{P\left(M_i \middle| \boldsymbol{X}\right)}{P\left(M_j \middle| \boldsymbol{X}\right)} \\
\iff & h_j\left(\boldsymbol{X} \middle| S(\boldsymbol{X})\right) = h_i\left(\boldsymbol{X} \middle| S(\boldsymbol{X})\right)
\end{aligned}
$$

A Fundamental Flaw of Bayes Factors.

- It can be shown that [3]:

$$
\begin{aligned}
B_{ij} &= \frac{P\left(M_i \middle| \boldsymbol{X}\right)}{P\left(M_j \middle| \boldsymbol{X}\right)} \times \frac{h_j\left(\boldsymbol{X} \middle| S(\boldsymbol{X})\right)}{h_i\left(\boldsymbol{X} \middle| S(\boldsymbol{X})\right)} \\
&= \frac{P\left(M_i \middle| \boldsymbol{X}\right)}{P\left(M_j \middle| \boldsymbol{X}\right)} \\
\iff & h_j\left(\boldsymbol{X} \middle| S(\boldsymbol{X})\right) = h_i\left(\boldsymbol{X} \middle| S(\boldsymbol{X})\right)
\end{aligned}
$$

- That is, B_{ij} will be biased unless the probability of seeing the data, given the observed summary statistics, is equal for each model.

Post-Hoc Model Comparison.

- Consider other problems with B_{ij} (and any post-hoc model comparison method).

Post-Hoc Model Comparison.

- Consider other problems with B_{ij} (and any post-hoc model comparison method).
- Posterior distributions are sensitive to choices of prior distributions.

Post-Hoc Model Comparison.

- Consider other problems with B_{ij} (and any post-hoc model comparison method).
- Posterior distributions are sensitive to choices of prior distributions.
- A particularly poor choice of $\pi_j(\theta)$ may reduce the number of retained simulations under Model j, and hence inflate B_{ij}.

Post-Hoc Model Comparison.

- We would like a model selection algorithm that avoids comparing posterior distributions.

Post-Hoc Model Comparison.

- We would like a model selection algorithm that avoids comparing posterior distributions.
- Given that our ‘semi-automatic summary selection’ version ABC is an example of ‘supervised learning’, we could consider a similar method for model selection.

Multiple Logistic Regression.

- Let $\boldsymbol{X}$ be our data (the collection of $\Gamma \times T$ summary statistics),

Multiple Logistic Regression.

- Let $\boldsymbol{X}$ be our data (the collection of $\Gamma \times T$ summary statistics),
- Let $\boldsymbol{x}^m = \left(s_1^m, \cdots, s_T^m\right)$ be the m^{th} row of $\boldsymbol{X}$ (the summary statistics from the m^{th} simulation).

Multiple Logistic Regression.

- Let $\boldsymbol{X}$ be our data (the collection of $\Gamma \times T$ summary statistics),
- Let $\boldsymbol{x}^m = \left(s_1^m, \cdots, s_T^m\right)$ be the m^{th} row of $\boldsymbol{X}$ (the summary statistics from the m^{th} simulation).
- Let Y^m be the category of the m^{th} observation (the model used for the m^{th} simulation).

Multiple Logistic Regression.

- Let $\boldsymbol{X}$ be our data (the collection of $\Gamma \times T$ summary statistics),
- Let $\boldsymbol{x}^m = \left(s_1^m, \cdots, s_T^m\right)$ be the m^{th} row of $\boldsymbol{X}$ (the summary statistics from the m^{th} simulation).
- Let Y^m be the category of the m^{th} observation (the model used for the m^{th} simulation).
- Let $\boldsymbol{\beta}^c = \left(\beta_0^c, \cdots, \beta_T^c\right)$ be the vector of coefficients for category c.

Multiple Logistic Regression.

- Let $\boldsymbol{X}$ be our data (the collection of $\Gamma \times T$ summary statistics),
- Let $\boldsymbol{x}^m = \left(s_1^m, \cdots, s_T^m\right)$ be the m^{th} row of $\boldsymbol{X}$ (the summary statistics from the m^{th} simulation).
- Let Y^m be the category of the m^{th} observation (the model used for the m^{th} simulation).
- Let $\boldsymbol{\beta}^c = \left(\beta_0^c, \cdots, \beta_T^c\right)$ be the vector of coefficients for category c.
- We aim to best fit the model

$$\ln\left(\frac{P(Y^m = c|\boldsymbol{X})}{P(Y^m = q|\boldsymbol{X})}\right) = \boldsymbol{\beta}^c \cdot \boldsymbol{x}',$$

for $c = 1, \cdots, J-1$.

Multiple Logistic Regression.

- We end up with a predictive model such that we can predict for $\boldsymbol{X}_{NEW}$:

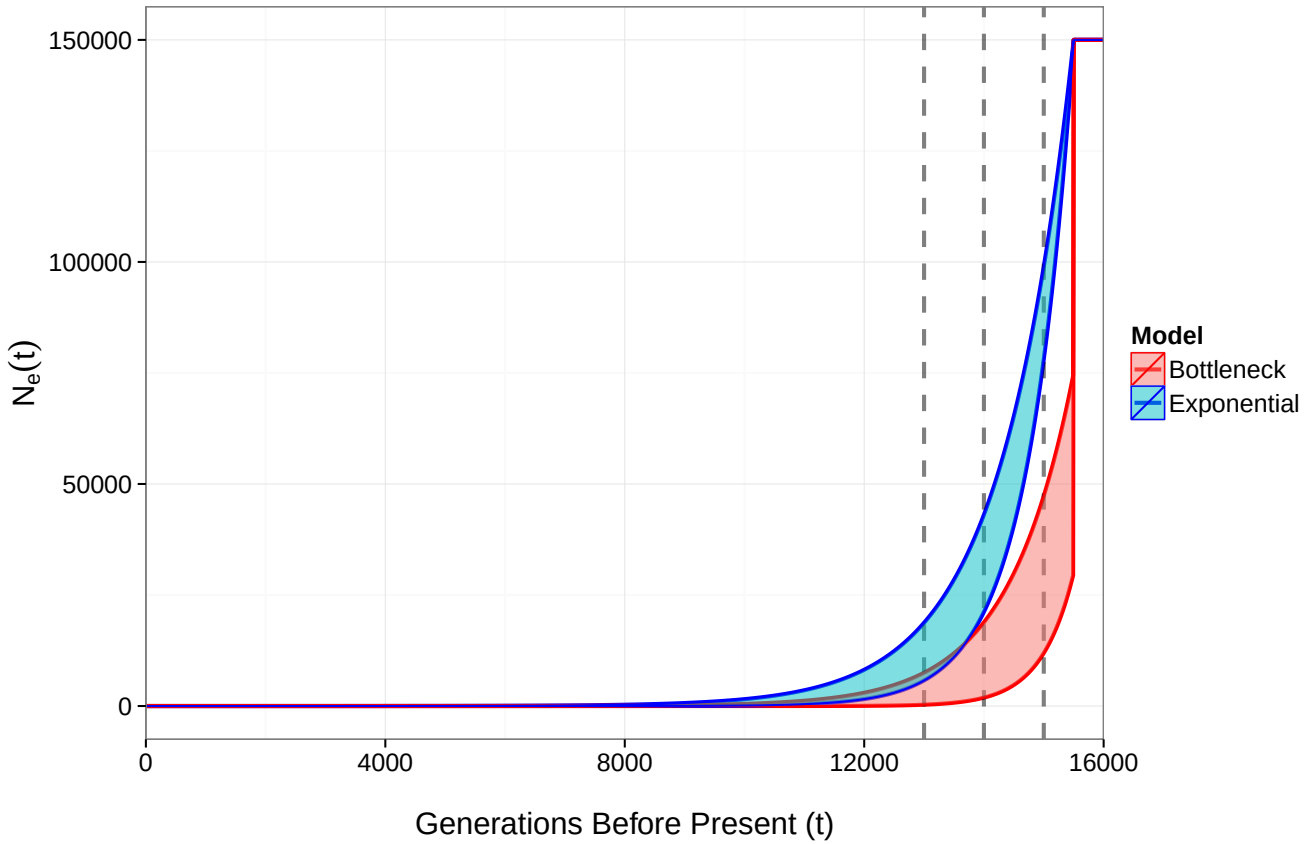
$$P(Y^m = c \big| \boldsymbol{X}_{NEW}) = p_c$$

for each $c \in \{1, \cdots, q\}$, such that

$$\sum_{i=1}^{q} p_i = 1.$$

Multiple Logistic Regression Example.

- Consider two opposing models of population dynamics:



Multiple Logistic Regression Example.

- The Bottleneck Model:
  - A sudden reduction to between 20% and 40% of the effective population size occurs before the species dies out.
- The Exponential Model:
  - There was no sudden population size reduction, the species just died out (relatively) slowly over 3000 generations.

Multiple Logistic Regression Example.

- However, we don't know which model fits our data best.
- If the data came from the Bottleneck Model, my prior belief is that: $N(16000) = 150,000$, $N(15500) \sim U(30,000, 75,000)$ and $N(12000) \sim U(300, 12500)$.
- If the data came from the Exponential Model, my prior belief is that: $N(16000) = 150,000$, $N(15500) = 150,000$ and $N(12000) \sim U(300, 7500)$.

Multiple Logistic Regression Example.

- I produced training data of this form with only 10,000 (5000 simulations from each model $\approx$ 2 mins), and fit the MLR (call this trainDat).

Multiple Logistic Regression Example.

- I produced training data of this form with only 10,000 (5000 simulations from each model $\approx$ 2 mins), and fit the MLR (call this trainDat).
- I then produced another 10,000 independent simulations (call this testDat).

Multiple Logistic Regression Example.

- I produced training data of this form with only 10,000 (5000 simulations from each model $\approx$ 2 mins), and fit the MLR (call this trainDat).
- I then produced another 10,000 independent simulations (call this testDat).
- Finally, I used the MLR to find which model I would *predict* had produced each of the 'testDat' simulations.

Multiple Logistic Regression Example.

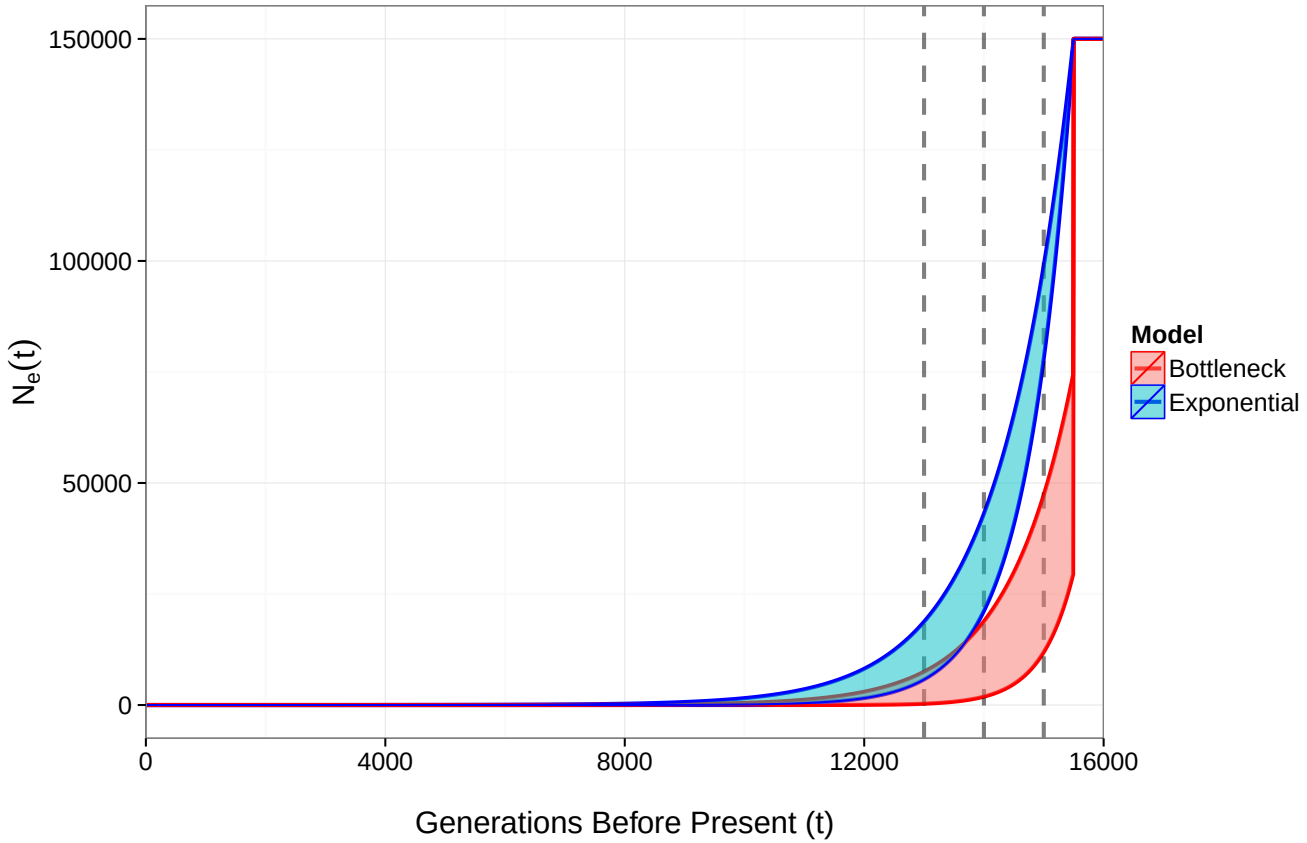
- I produced training data of this form with only 10,000 (5000 simulations from each model $\approx$ 2 mins), and fit the MLR (call this trainDat).
- I then produced another 10,000 independent simulations (call this testDat).
- Finally, I used the MLR to find which model I would *predict* had produced each of the 'testDat' simulations.
- The model predicted correctly for 99.53% of the testDat simulations (total 4.5 minutes).

Multiple Logistic Regression Example.

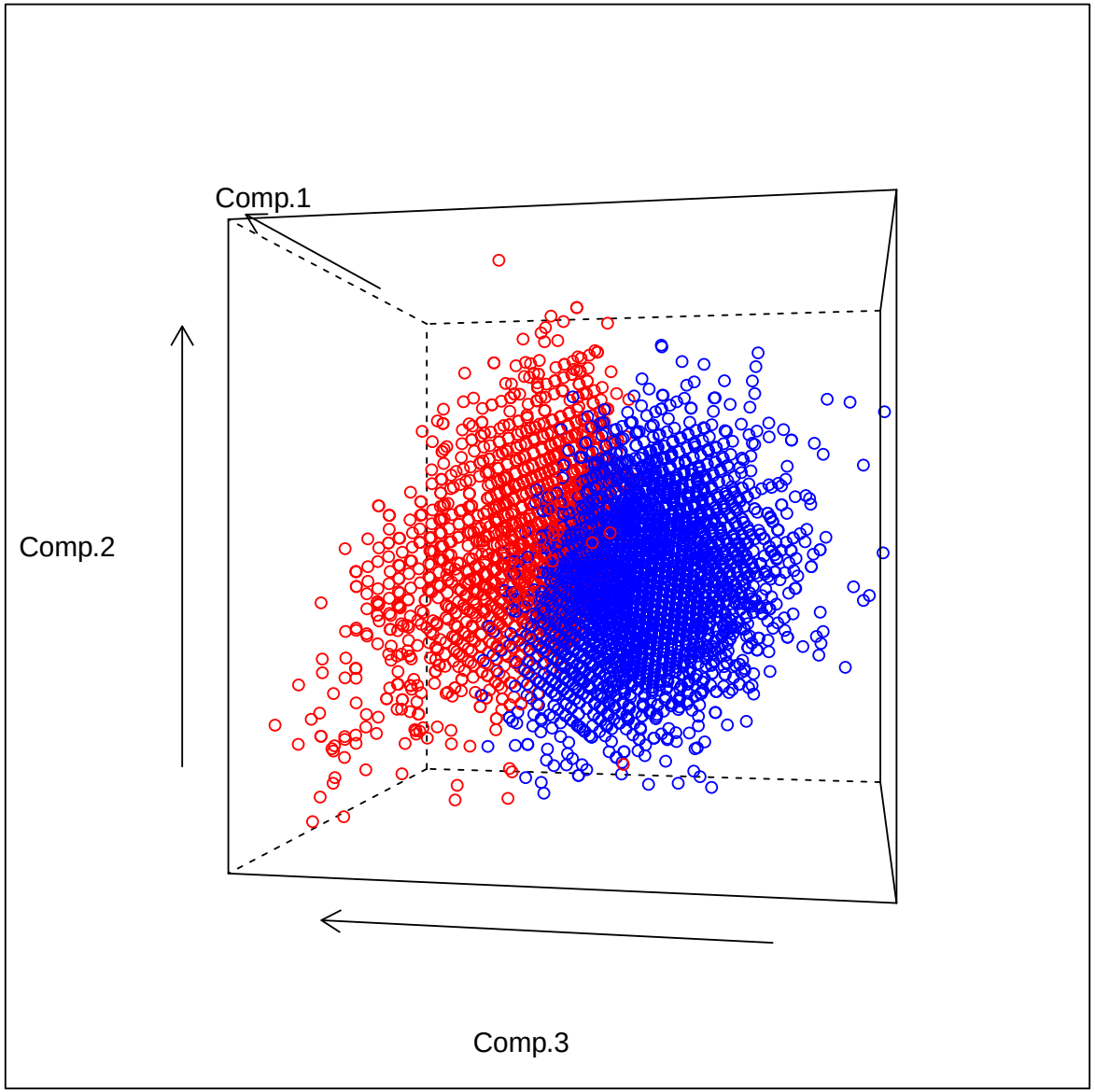
- I produced training data of this form with only 10,000 (5000 simulations from each model $\approx$ 2 mins), and fit the MLR (call this trainDat).
- I then produced another 10,000 independent simulations (call this testDat).
- Finally, I used the MLR to find which model I would *predict* had produced each of the 'testDat' simulations.
- The model predicted correctly for 99.53% of the testDat simulations (total 4.5 minutes).
- A corresponding Bayes Factor Analysis returned 17.03% accuracy (total 21 minutes).

Multiple Logistic Regression Example.

- Recall the two opposing models of population dynamics:



Multiple Logistic Regression Example.



Conclusions.

- In my thesis we performed a four model Semi-Automatic ABC Analysis.

Conclusions.

- In my thesis we performed a four model Semi-Automatic ABC Analysis.
- Our MLR classification returned $> 96\%$ accuracy for $> 250,000$ simulations.

Conclusions.

- In my thesis we performed a four model Semi-Automatic ABC Analysis.
- Our MLR classification returned $> 96\%$ accuracy for $> 250,000$ simulations.
- A complimentary Bayes Factor analysis never returned a correct post-hoc analysis for our simulated data.

THE UNIVERSITY of ADELAIDE

Conclusions.

- In my thesis we performed a four model Semi-Automatic ABC Analysis.
- Our MLR classification returned > 96% accuracy for $> 250,000$ simulations.
- A complimentary Bayes Factor analysis never returned a correct post-hoc analysis for our simulated data.
- Our method does not require ABC to be performed on all possible models (just simulations).

THE UNIVERSITY of ADELAIDE

Thanks.

- Dr Barbara Holland and Dr Jeremy Sumner.
- Prof. Nigel Bean and Dr Jono Tuke.
- Prof. Alan Cooper and everyone at ACAD.
- ACEMS for funding my visit.

THE UNIVERSITY of ADELAIDE

[1] M.A. Beaumont. Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41:379–406, 2010.

[2] P. Fearnhead and D. Prangle. Constructing Summary Statistics for Approximate Bayesian Computation: Semi-automatic Approximate Bayesian Computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, June 2012.

[3] C. Robert, J-M. Cornuet, J-M. Marin, and N.S. Pillai. Lack of Confidence in Approximate Bayesian Computation Model Choice. *Proceedings of the National Academy of Sciences*, 108(37):15112–15117, 2011. doi: 10.1073/pnas.1102900108. URL `http://www.pnas.org/content/108/37/15112.abstract`.

[4] B. Shapiro, A. J. Drummond, A. Rambaut, M. C. Wilson, P. E. Matheus, A. V. Sher, O. G. Pybus, M. T. P. Gilbert, I. Barnes, J. Binladen, E. Willerslev, A. J. Hansen, G. F. Baryshnikov, J. A. Burns, S. Davydov, J. C. Driver, D. G. Froese, C. R. Harington, G. Keddie, P. Kosintsev, M. L. Kunz, L. D. Martin, R. O. Stephenson, J. Storer, R. Tedford, S. Zimov, and A. Cooper. Rise and Fall of the Beringian Steppe Bison. *Science*, 306:1561–1565, November 2004.