Introduction

- USC
  Molecular and Computational Biology
  * Bioinformatics/genomics
  * CEGS – HapMap Project
- Cambridge
  Oncology
  * cancer genomics
  DAMTP
  * stochastic computation

Outline

- Four examples of evolutionary inference on different time scales
  - Origin of primates
  - Genome-wide patterns of SNP variation and effects on gene expression
  - Recombination rates across the genome
  - Cancer genomics
- Stochastic computation
  - Simulating posterior distributions
  - Approximate Bayesian Computation

Ex. 1. Primate Phylogeny

Ex. 2. Genome-wide patterns of SNP variation

Ex. 3. Recombination hotspots
Ex. 4. Cancer genomics, stem cells

Statistical aspects

- Highly dependent data
- Dependence caused by tree or graph linking observations (ancestral history)
- Explicit theory hard to come by . . .
- . . . computational inference methods essential

Outline of Lectures

- Stochastic computation
- Cancer genomics and stem cells
- Recombination hotspots
- SNPs and expression variants
- Inference in fossil record
- Random combinatorial structures

Stochastic computation in evolutionary genetics

- Many different approaches have been employed:
  - Rejection
  - Importance sampling
  - Sequential importance sampling with resampling
  - Markov chain Monte Carlo
  - Metropolis-coupled MCMC
  - Perfect sampling: coupling from the past

Stochastic computation in evolutionary genetics

- Many different approaches have been employed:
  - Rejection
  - Importance sampling
  - Sequential importance sampling with resampling
  - Markov chain Monte Carlo
  - Metropolis-coupled MCMC
  - Perfect sampling: coupling from the past

Caveat: described as separate approaches, but often combined!

Introduction to Bayesian computation

- Discrete data \( \mathcal{D} \), prior \( \pi(\theta) \) for parameters \( \theta \)
- Want to generate observations from posterior distribution \( f(\theta | \mathcal{D}) \)
- Bayes Theorem gives

\[
f(\theta | \mathcal{D}) = \frac{\mathbb{P}(\mathcal{D} | \theta)\pi(\theta)}{\mathbb{P}(\mathcal{D})}, \quad (1)
\]

where the normalizing constant is

\[
\mathbb{P}(\mathcal{D}) = \int \mathbb{P}(\mathcal{D} | \tau)\pi(\tau) \, d\tau
\]

Posterior proportional to likelihood \( \times \) prior
Rejection methods I

R1 Generate $\theta$ from $\pi(\cdot)$

R2 Accept $\theta$ with probability $\mathbb{P}(\mathcal{D} \mid \theta)$; return to [R1]

Accepted observations have distribution $f(\theta \mid \mathcal{D})$

Proof: the density of accepted points is just the prior density times the rate at which such points are accepted:

$$f(\theta \mid \mathcal{D}) \propto \pi(\theta) \times \mathbb{P}(\mathcal{D} \mid \theta),$$

which, from (1), does the trick.

An improvement

Note that if you can find a constant $c$ such that $\mathbb{P}(\mathcal{D} \mid \theta) \leq c$, for all $\theta$,

then we can replace [R2] above with

[R2’ ] Accept $\theta$ with probability $\mathbb{P}(\mathcal{D} \mid \theta)/c$; return to [R1]

The best choice of $c$ can be found from the maximum likelihood estimator of $\theta$:

$$c = \max_\tau \mathbb{P}(\mathcal{D} \mid \tau) \equiv \mathbb{P}(\mathcal{D} \mid \hat{\theta}).$$

The normalizing constant

To evaluate the behavior of the algorithm, we calculate the distribution of the number $R$ of steps up to and including an accepted observation. Write

$$p(\theta) = \frac{\mathbb{P}(\mathcal{D} \mid \theta)}{c}$$

and note that if $\theta \sim \pi(\cdot)$, then

$$p \equiv \mathbb{E}p(\theta) = \int \frac{\mathbb{P}(\mathcal{D} \mid \theta)}{c} \pi(\theta) d\theta = \frac{\mathbb{P}(\mathcal{D})}{c}$$

The number of runs to get $n$ observations from $f(\theta \mid \mathcal{D})$ is therefore negative binomial, with mean

$$\frac{n}{p} = \frac{nc}{\mathbb{P}(\mathcal{D})}$$

This shows

- the effect of $\mathbb{P}(\mathcal{D})$
- the effect of $c$
- the way to estimate $\mathbb{P}(\mathcal{D})$ (Bayes factors)

Complex stochastic models

- A stochastic process often underlies the likelihood computation
- This process may be complex, making explicit probability calculations difficult or impossible
- Thus $\mathbb{P}(\mathcal{D} \mid \theta)$ may be uncomputable (either quickly or theoretically)

When the stochastic process is easy to simulate . . .
Rejection methods II

RS1 Generate $\theta$ from $\pi(\cdot)$
RS2 Simulate $D'$ from stochastic model with parameter $\theta$
RS3 Accept $\theta$ if $D' = D$; return to 1

Just as before, accepted observations from this algorithm have the density $f(\cdot|D)$
Despite its appearance, this algorithm is much more general than our first one, precisely because we do not have to compute likelihoods. We “just” need to simulate the process.

Approximate Bayesian Computation I

A1 Generate $\theta$ from $\pi(\cdot)$
A2 Simulate $D'$ from stochastic model with parameter $\theta$
A3 Calculate distance $\rho(D, D')$ between $D'$ and $D$
A4 Accept $\theta$ if $\rho \leq \epsilon$; return to 1

- If $\epsilon \to \infty$, generates from prior
- If $\epsilon = 0$, generates from $f(\theta | D)$
- Choice of $\epsilon$ reflects tension between computability and accuracy
  - PCR — post-computational remorse
- Method is honest: you get observations from $f(\theta | \rho(D, D') \leq \epsilon)$

Approximate Bayesian Computation II

- The limit $\epsilon = 0$ reproduces the data precisely
- In many examples the data are too high-dimensional
  - ... so reduce dimension by using summary statistics

I am not aware of a trick analogous to normalizing by $c$ to speed up this method.

- The simulation step raises the question of whether we ever hit the target
- The next approach is the simplest ABC method
Motivation: sufficiency

Recall that \( S = S(D) \) is sufficient for \( \theta \) if
\[
\mathbb{P}(D \mid S, \theta) \text{ is independent of } \theta
\]
- If \( S \) is sufficient for \( \theta \), then
\[
f(\theta \mid D) = f(\theta \mid S)
\]

\[
f(\theta \mid D) \propto \mathbb{P}(D \mid \theta) \pi(\theta)
= \sum_s \mathbb{P}(D \mid S = s, \theta) \mathbb{P}(S = s \mid \theta) \pi(\theta)
= \text{constant} \times \mathbb{P}(S \mid \theta) \pi(\theta)
\propto f(\theta \mid s)
\]

• Typically, \( S \) is of smaller dimension that \( D \). Inference method can be simplified (and sped up), e.g. rejection method via
\[
f(\theta \mid S) \propto \mathbb{P}(S \mid \theta) \pi(\theta)
\]

• Example: estimating \( \theta \) from the Ewens Sampling Formula. We saw that the number of types in the sample was sufficient for \( \theta \)

Research problem

Puts a premium on finding decent summary statistics

- Definition of approximate sufficiency? (LeCam 1963)
- A systematic, implementable approach?
- Estimate distance between \( f(\theta \mid D) \) and \( f(\theta \mid S) \) given a measure of how far from sufficient \( S \) is for \( \theta \)

Combine summaries and rejection


Choose statistics \( S = (S_1, \ldots, S_p) \) to summarize \( D \)

AS1 Generate \( \theta \) from \( \pi(\cdot) \)
AS2 Simulate \( D' \), calculate \( s' \)
AS3 Accept \( \theta \) if \( \rho(s', s) \leq \epsilon \); return to 1

An example from population genetics

We have seen that the coalescent (or ARG) is a useful tool for interpreting patterns of molecular variation

Ingredients:
- Biological model for relationships among \( n \) sampled sequences (tree or graph; \( \Lambda, T \))
- Parameters of interest (mutation, migration, recombination, \ldots rates \( \theta \))
- Aim: generate observations from \( f(\theta \mid D) \) or from \( f(\theta, \Lambda, T \mid D) \)

Ward’s mtDNA data

Yakima \( n = 42 \), 360 bp region of mtDNA

Summary statistic is \( S \), the number of segregating sites.
We observed \( S = 31 \)

Interested (for now) in inference about
- the scaled mutation parameter \( \theta \) (no recombination here)
- the TMRCA \( T \) of sample ( = height of tree)
- Recall that \( T \) has prior mean 1.95

Now run rejection algorithm as follows:
Data augmentation

Write $G$ for the coalescent tree (topology and branch lengths). Then

$$f(\theta, G|S = s) = \mathbb{P}(S = s|G, \theta)\pi(G, \theta)$$

$$= \mathbb{P}(S = s|G, \theta)\pi(G)\pi(\theta)$$

- The prior for $G$ comes from the coalescent model (“simulate a random coalescent tree”)
- The likelihood term can be calculated by noting that if we know $G$, then we know the total length $L = \ell$ of the branches.

Results (wide flat prior for $\theta$)

<table>
<thead>
<tr>
<th>Acceptance rate</th>
<th>0.6%</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TMRCA$ $T$</td>
<td></td>
</tr>
<tr>
<td>1st quartile</td>
<td>1.07</td>
</tr>
<tr>
<td>mean</td>
<td>1.71</td>
</tr>
<tr>
<td>median</td>
<td>1.49</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>2.11</td>
</tr>
</tbody>
</table>

Given $\ell$ and $\theta$, we have:

$$h := \mathbb{P}(S = s|G, \theta) = \mathbb{P}(\text{Po}(\theta\ell/2) = s) = e^{-\theta\ell/2}(\theta\ell/2)^s/s!$$

- Simulate $\theta$ from $\pi$ and $G$ from the coalescent
- Compute the total length $\ell$ of the edges of $G$
- Keep $\theta, G$ with probability $h$, and repeat
- In this case, the magic constant $c$ is

$$c = \mathbb{P}(\text{Po}(s) = s) = s^s e^{-s}/s!$$

... so replace $h$ by $h/c$

ABC III: Generalization of ABC II

Beaumont, Zhang and Balding (2002), Genetics

A method that makes use of observations in a better way

Start by simulating pairs $(\theta_i, s_i), i = 1, 2, \ldots, m$

For motivation, suppose we could represent $f(\theta|S = s)$ as a regression model,

$$\theta_i = \alpha + (s_i - s)^\prime \beta + \epsilon_i, \ i = 1, 2, \ldots, m,$$

where the $\epsilon_i$ are uncorrelated, 0 mean, common variance.

When $s_i = s$, the observations come from the posterior with mean $\mathbb{E}(\theta|S = s) = \alpha$.

Local-linear regression

Linearity and additivity assumptions above are often implausible, but may apply locally in neighborhood of $s$

So use local-linear regression, by minimizing

$$\sum_{i=1}^m [\theta_i - \alpha - (s_i - s)^\prime \beta]^2 K_\delta(||s_i - s||),$$

where $K_\delta(\cdot)$ is a kernel function. They use

$$K_\delta(t) = \begin{cases} c\delta^{-1}(1 - (t/\delta)^2), & t \leq \delta \\ 0, & t > \delta \end{cases}$$

where $c$ is a normalizing constant.
Posterior means

The posterior mean estimate is

$$\hat{\alpha} = \frac{\sum_i \theta^*_i K_\delta(||s_i - s||)}{\sum_i K_\delta(||s_i - s||)}$$

For local-constant regression (i.e. $\beta = 0$) and indicator kernel

$$I_\delta(t) = \begin{cases} 1, & t \leq \delta \\ 0, & t > \delta \end{cases}$$

we get

$$\hat{\alpha} = \frac{\sum_i \theta_i I_\delta(||s_i - s||)}{\sum_i I_\delta(||s_i - s||)}$$

which is the usual rejection estimator (the weights are uniform over the number of accepted observations)

The role of $\delta$

- Choice of metric: usually scale the $s_i$ values to have equal variances, and take $||s||$ to be Euclidean norm (so acceptance regions are spheres)
- Choice of tolerance: for both rejection and regression methods, they set $\delta$ to be a quantile $p_\delta$ of the empirical distribution of the simulated $||s_i - s||$
- If $\theta$ is multi-dimensional, then $\beta$ is a matrix
- Routines are implemented in R

Advantages and disadvantages of ABC

Pros:

- Often easy to code
- Generates independent observations (can use embarrassingly parallel computation)
- Can be used to estimate Bayes factors directly
- Usually easy to adapt

Cons:

- May be hard to anticipate effects of summary statistics
- For complex probability models, sampling from prior does not make good use of accepted observations
- Choice of metric can matter
- Homework: find a scheme that combines generation of perfect observations with use of existing sample

Markov chain Monte Carlo

One of the most useful tricks in the statistical computing arsenal. It provides a way of simulating from a density $f$ known only up to a normalizing constant.

Although it is quite old (Metropolis et al. (1953), Hastings (1970)), it did not find its way into the statistics literature until relatively recently.

The idea is to generate observations on a Markov chain $X(t), t = 1, 2, \ldots$ which has the desired density as its stationary distribution.

If this chain is ergodic, we can estimate properties of $f$ from the simulated values.
There is now an extensive literature on this method. Useful places to start:


Liu JS (2001) *Monte Carlo strategies in scientific computing* Springer


The Hastings Markov chain

M1 Now at θ
M2 Propose move to θ’ according to q(θ → θ’)
M3 Calculate the Hastings ratio

\[ h = \min \left( 1, \frac{\mathbb{P}(D \mid \theta')\pi(\theta')q(\theta' \to \theta)}{\mathbb{P}(D \mid \theta)\pi(\theta)q(\theta \to \theta')} \right) \]
M4 Accept θ’ with probability h, else return θ

Checking it works

To check that \( f(\theta|D) \) is indeed the stationary distribution, we check the detailed balance equations. If \( r(\theta, \theta') \) denotes the transition mechanism of the chain, then we show that

\[ f(\theta|D)r(\theta, \theta') = f(\theta'|D)r(\theta', \theta) \]

Some special cases

There are several special cases of this, including

- Independence sampler: \( q(\theta, \theta') = q(\theta') \). \( h \) depends on ratios of likelihoods, priors and proposals
- Metropolis sampler: \( q(\theta, \theta') = q(\theta', \theta) \). \( h \) depends on likelihood and ratio of priors
- Reversible sampler:

\[ \pi(\theta)q(\theta, \theta') = \pi(\theta')q(\theta', \theta) \]

In this case, \( h \) depends on likelihood ratio

Basic output analysis

There are more things to check:

- Is the chain ergodic?
- Does it mix well?
- Is the chain stationary?
- Have we burnt in enough?
- Diagnostics of the run (e.g. CODA)

A toy example

To illustrate, we consider the following example

- \( X \sim \text{Poi}(\theta) \)
- \( \theta \sim U[0, T] \)

In this case the posterior is a truncated gamma:

\[ f(\theta|X = x) \propto e^{-\theta}x^x, \quad 0 < \theta < T \]
Suppose we observe $X = 12$, and take $T = 20$.

For the update rule:

- generate $U \sim U(0, 1)$
- set $V = \epsilon(U - 0.5)$
- set $\theta' = \theta + V$, folded if it goes outside $(0, T)$
- can check this is a reversible case (so only need likelihood ratio)
Simulated and exact density

Histogram of theta values

MCMC in evolutionary genetics setting

- Small tweaks in the biology often translate into huge changes in algorithm
- Long development time
- All the usual problems with convergence
- Almost all the effort goes into evaluation of likelihood

(Yet) another MCMC approach

MS1 Now at $\theta$

MS2 Propose a move to $\theta'$ according to $q(\theta \rightarrow \theta')$

MS3 Generate $\mathcal{D}'$ using $\theta'$

MS4 If $\mathcal{D}' = \mathcal{D}$, go to next step, else return $\theta$

MS5 Calculate

$$h = h(\theta, \theta') = \min \left( 1, \frac{\pi(\theta')q(\theta' \rightarrow \theta)}{\pi(\theta)q(\theta \rightarrow \theta')} \right)$$

MS6 Accept $\theta'$ with probability $h$, else return $\theta$

Theorem

The stationary distribution of the no-likelihood chain is $f(\theta | \mathcal{D})$

Proof

Denote the transition mechanism of the chain by $r(\theta \rightarrow \theta')$

Choose $\theta' \neq \theta$ satisfying

$$\frac{\pi(\theta')q(\theta' \rightarrow \theta)}{\pi(\theta)q(\theta \rightarrow \theta')} \leq 1$$

Return to our Poisson example

We run the no-likelihood method with the same update as the MCMC case.

The simulations assume $\epsilon = 1$, so we accept observations within 1 of $X = 12$
Practical version: ABC

Data $\mathcal{D}$, summary statistics $S$

[MS4'] If $\rho(\mathcal{D}', \mathcal{D}) \leq \epsilon$, go to next step, otherwise return $\theta$

[MS4''] If $\rho(S', S) \leq \epsilon$, go to next step, otherwise return $\theta$

for some suitable metric $\rho$ and approximation level $\epsilon$

Observations now from $f(\theta | \rho(\mathcal{D}', \mathcal{D}) \leq \epsilon)$ or $f(\theta | \rho(S', S) \leq \epsilon)$

Variations on ABC

- These methods can often be started at stationarity, so no burn-in
- If the underlying probability model is complex, simulating data will not often lead to acceptance. Thus need update for parts of the probability model (data augmentation)
- What if $\mathcal{D}$ is not discrete?
  - Use previous method (binning)
  - Use simulation approach to estimate the likelihood terms in the Hastings ratio (Diggle & Gratton, RSSB, 1980)

Another tweak: augmenting state space

MS1 Now at $(\theta, T)$

MS2 Propose a move to $(\theta', T')$ according to $q((\theta, T) \rightarrow (\theta', T'))$

MS3 Simulate $\mathcal{D}'$ using $(\theta', T')$

MS4 If $\mathcal{D}' = \mathcal{D}$, go to next step, else return $(\theta, T)$ (or use ABC)

MS5 Calculate

$$h = \min \left( 1, \frac{\pi(\theta', T')q((\theta', T') \rightarrow (\theta, T))}{\pi(\theta, T)q((\theta, T) \rightarrow (\theta', T'))} \right)$$

MS6 Accept $(\theta', T')$ with probability $h$, else return $(\theta, T)$.

Mitochondrial example revisited

$H =$ number of haplotypes

$V =$ number of SNPs

- Yakima $n = 42$, $V = 31$, $H = 20$
- Nuu Chah Nulth $n = 63$, $V = 26$, $H = 28$

Yakima results

<table>
<thead>
<tr>
<th>$S = V, \epsilon = 0$</th>
<th>Rej.</th>
<th>Estimated like.</th>
<th>No like.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptance rate</td>
<td>0.6%</td>
<td>65%</td>
<td>3.9%</td>
</tr>
<tr>
<td>$TMRCA T$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1st quartile</td>
<td>1.07</td>
<td>1.05</td>
<td>1.09</td>
</tr>
<tr>
<td>mean</td>
<td>1.71</td>
<td>1.70</td>
<td>1.72</td>
</tr>
<tr>
<td>median</td>
<td>1.49</td>
<td>1.45</td>
<td>1.49</td>
</tr>
<tr>
<td>3rd quartile</td>
<td>2.11</td>
<td>2.07</td>
<td>2.15</td>
</tr>
</tbody>
</table>

Interested in inference about

- the scaled mutation parameter $\theta$
- the TMRCA $T$ of sample ( = height of tree)
- $T$ has prior mean 1.95

From MCMC methods

Yakima $\mathbb{E}(T | \mathcal{D}) = 0.72$

Nuu Chah Nulth $\mathbb{E}(T | \mathcal{D}) = 0.68$
\[ S = (V, H), \epsilon = 0 \]

<table>
<thead>
<tr>
<th>Acceptance rate</th>
<th>Rejection</th>
<th>No likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.03%</td>
<td>0.46%</td>
</tr>
</tbody>
</table>

**TMRCA**

| 1st quartile | 0.73 | 0.74 |
| mean         | 1.01 | 1.03 |
| median       | 0.93 | 0.94 |
| 3rd quartile | 1.22 | 1.24 |

**The effects of \( \epsilon \)**

\[ S = (V, H) \]

<table>
<thead>
<tr>
<th>( \epsilon )</th>
<th>0.2%</th>
<th>0.04%</th>
<th>0.005%</th>
</tr>
</thead>
</table>

**TMRCA**

| 1st quartile | 0.54 | 0.49 | 0.46 |
| mean         | 0.70 | 0.64 | 0.59 |
| median       | 0.66 | 0.60 | 0.55 |
| 3rd quartile | 0.81 | 0.74 | 0.69 |

**Summary**

In the population genetics setting, there are many existing MCMC algorithms that explore the space of \( \theta, \Lambda, T \).

These can be used immediately to address harder problems, by replacing likelihood computations by simulation.

- Complex mutation models (e.g. dependent sites in a sequence)
- RFLPs: Population expansion, multiple cutters, mutation model
- Ascertainment (e.g. SNP panels)

**Current work**

- Exploring connection with other auxiliary variable methods
  - Pettitt and Reeves — unified treatment that includes indirect inference
  - Sisson et al. — variety of tricks for speeding up mixing in no-likelihood methods. No-likelihood sequential importance sampling
- Exploring projection pursuit for combining summary statistics [Lecture 3]
- Exploring Gibbs sampler with no likelihoods [Lecture 5]