Recap

- Last time, we described some of the mathematical underpinnings from population genetics, showed how they had influenced the field, and what they had led to.
- The Ewens Sampling Formula paper (1972) started the field of inference in population genetics, and has many other uses in combinatorics and Bayesian non-parametric analysis.
- We described an example from the world of computational cancer genomics, in which ABC could be used to attack what is in essence a combinatorics problem.
- We alluded to the need for faster coalescent simulation approaches.

Regression-based methods

Motivation – 1

The idea is to replace the hard cut-off in the ABC with a soft version that makes use of all of the observations.

In the summary statistic approach care has to be taken to choose $\rho(S, S')$.

For example, if $S = (S_1, \ldots, S_m)$ is an $m$-dimensional summary, and

$$\rho(S, S') = ||S' - S|| = \sqrt{\sum_{i=1}^{m} (S'_i - S_i)^2}$$

then accepting whenever $\rho \leq \epsilon$ treats the values of $S'$ equally, regardless of the value of $\rho$.

- smooth weighting
- regression adjustment

The method is insensitive to the value of $\epsilon$, and allows $m$ to be large.
**Motivation – 2**

We start from $M$ observations $(\theta_i, S_i)$, where each $\theta_i$ is an independent draw from the prior $\pi(\cdot)$ and $S_i$ is the set of summary statistics generated when $\theta = \theta_i$.

We standardize the coordinates of $S_i$ to have equal variances.

Now the posterior is

$$f(\theta|S) = \frac{f(\theta, S)}{f(S)}$$

so to estimate the left-hand side we could estimate the joint density and the marginal likelihood, and evaluate at $S = s_0$, the data.

---

**Motivation – 3**

The $(\theta_i, S_i)$ are a sample from the joint law, and the rejection method is just one way to estimate the conditional law when $S = s_0$; those with small values of $||S_i - s_0||$ are the ones to use.

This can be improved by

- weighting the $\theta_i$ according to $\rho(S_i, s_0)$
- adjusting the $\theta_i$ by local-linear regression

Imagine that we have

$$\theta_i = m(s_i) + \epsilon := \alpha + (s_i - s_0)^T \beta + \epsilon_i, i = 1, 2, \ldots, M$$

(3)

where the $\epsilon_i$ are uncorrelated $(0, \sigma^2)$ rvs
Motivation – 4

When \( s_i = s_0 \), \( \theta_i \) is drawn from a distribution with mean

\[
\mathbb{E}(\theta|S = s_0) = \alpha
\]

The least-squares estimator of \((\alpha, \beta)\) minimizes

\[
\sum_{i=1}^{M} (\theta_i - \alpha - (s_i - s_0)^T \beta)^2
\]

so that

\[
(\hat{\alpha}, \hat{\beta}) = (X^T X)^{-1} X^T \theta,
\]

where \( X \) is the design matrix.

Motivation – 5

\[
X = \begin{pmatrix}
1 & s_{11} - s_{01} & \cdots & s_{1m} - s_{0m} \\
\vdots & \vdots & & \vdots \\
1 & s_{M1} - s_{01} & \cdots & s_{Mm} - s_{0m}
\end{pmatrix}
\]

Then, from (3)

\[
\theta_i^* = \theta_i - (s_i - s_0)^T \hat{\beta}
\]

form an approximate random sample from \( f(\theta|s_0) \)

Note that \( \mathbb{E}(\theta|s_0) = \hat{\alpha} = M^{-1} \sum \theta_i^* \)
Regression method – 1

We can improve things by using weighted regression. Replace the minimization objective with

$$\sum_{i=1}^{M}(\theta_i - \alpha - (s_i - s_0)^T \beta)^2 K_\epsilon(||s_i - s_0||)$$ (4)

One choice is the Epanechnikov kernel

$$K_\epsilon(t) = \frac{3}{2\epsilon} \left( 1 - \left( \frac{t}{\epsilon} \right)^2 \right) \mathbb{I}(t \leq \epsilon);$$

$$\int_0^\epsilon K_\epsilon(t) dt = 1.$$ Now we get

$$(\hat{\alpha}, \hat{\beta}) = (X^TWX)^{-1}X^TW\theta,$$

Regression method – 2

where $W = \text{diag}\{K_\epsilon(||s_i - s_0||)\}$

We now have

$$\mathbb{E}(\theta | s_0) = \hat{\alpha} = \frac{\sum \theta_i^* K_\epsilon(||s_i - s_0||)}{\sum K_\epsilon(||s_i - s_0||)}$$

Our weighted sample from the posterior is given by $(\theta_i^*, w_i)$, with

$$w_i = \frac{K_\epsilon(||s_i - s_0||)}{\sum K_\epsilon(||s_i - s_0||)}, i = 1, \ldots, M$$

This discussion concerns linear adjustment. Can generalize to other regression models by replacing $\alpha + (s_i - s_0)^T \beta$ in (4) by an appropriate $m(s_i)$. 

73
Regression method – 3: Special cases

- For local-constant regression, we set $\beta = 0$
- Then if $K_\epsilon(\cdot)$ is replaced by
  \[ I_\epsilon(t) = \epsilon^{-1} \mathbb{I}(t \leq \epsilon) \]
  get
  \[ \hat{\alpha} = \frac{\sum \theta_i I_\epsilon(||s_i - s_0||)}{\sum I_\epsilon(||s_i - s_0||)} \]
  which is the rejection method estimate

Choice of $\epsilon$

- Set $\epsilon$ to be a quantile, $P_\epsilon$, of the empirical distribution of simulated values of $||s_i - s_0||$
- The choice of $\epsilon$ involves, as ever, a trade-off between bias and variance:
  - As $\epsilon \uparrow$, you use more observations, so less variance . . .
  - . . . but more bias

Note: as $\epsilon \downarrow 0$, both rejection and regression methods are equivalent

Implementation: the abc package in R implements these and many other methods.
Heteroscedastic regression

There is a literature on heteroscedastic models. The regression equation is

\[ \theta = m(s) + \sigma(s)\xi \]

where \( \sigma(s) \) is the square root of conditional variance of \( \theta \) given \( s \), and \( \xi \) is the residual. Using \( \hat{\cdot} \) to denote estimator, the adjusted observations are

\[ \theta^*_i = \hat{m}(s_0) + \hat{\sigma}(s_0)\hat{\xi}_i = \hat{m}(s_0) + \frac{\hat{\sigma}(s_0)}{\hat{\sigma}(s_i)}(\theta_i - \hat{m}(s_i)) \]


Markov Chain Monte Carlo methods

MCMC – 1

The idea is to construct an ergodic Markov chain that has \( f(\theta|D) \) as its stationary distribution, in the case that normalising constants cannot be computed. Here is Hastings’ (Biometrika, 1970) classic method:

1. Now at \( \theta \)
2. Propose move to \( \theta' \) according to \( q(\theta \rightarrow \theta') \)
3. Calculate the Hastings ratio

\[ h = \min \left(1, \frac{\mathbb{P}(D \mid \theta')\pi(\theta')q(\theta' \rightarrow \theta)}{\mathbb{P}(D \mid \theta)\pi(\theta)q(\theta \rightarrow \theta')} \right) \]

4. Accept \( \theta' \) with probability \( h \), else return \( \theta \)
MCMC – 2

There are more things to check:
• Is the chain ergodic?
• Does it mix well?
• Is the chain stationary?
• Burn in?
• Diagnostics of the run (no free lunches) – see coda package in R for example

MCMC – 3

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
Here is an ABC version (Marjoram et al, PNAS, 2003)

1. Now at $\theta$
2. Propose a move to $\theta'$ according to $q(\theta \rightarrow \theta')$
3. Generate $D'$ using $\theta'$
4. If $D' = D$, go to next step, else return $\theta$
5. Calculate
   $$ h = h(\theta, \theta') = \min \left( 1, \frac{\pi(\theta')q(\theta' \rightarrow \theta)}{\pi(\theta)q(\theta' \rightarrow \theta')} \right) $$
6. Accept $\theta'$ with probability $h$, else return $\theta$

Lemma: The stationary distribution of the chain is, indeed, $f(\theta|D)$.

Proof: In class . . .
ABC-MCMC – 3

Here is the practical version, for data $\mathcal{D}$, summary statistics $S$

4'. If $\rho(D', D) \leq \epsilon$, go to next step, otherwise return $\theta$

4''. 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 $f(\theta \mid \rho(D', D) \leq \epsilon)$ or $f(\theta \mid \rho(S', S) \leq \epsilon)$

Variations on a theme – 1

There have been many variants on the theme. For example, one might use multiple simulations from a given $\theta$ to get a better estimate of the likelihood. This is known as the pseudo-marginal method (Beaumont, *Genetics*, 2003; Tavaré et al. *PNAS*, 2003; Andrieu & Roberts *Ann Statist*, 2009). The idea is to simulate pairs of data points $(\theta, \hat{P}(D \mid \theta))$: 

2'. If at $\theta'$ simulate $B$ values of $D'$, and use these to estimate $P(D \mid \theta')$ via

$$\hat{P}(D \mid \theta') = \frac{1}{B} \sum_{j=1}^{B} 1_{(D' = D)}$$

3'. If this is 0, stay at $\theta$; else
Variations on a theme – 2

4’. Accept $\theta'$ and $\hat{P}(D|\theta')$ with probability

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

else stay at $\theta$.

Variations on a theme – 3

- Convergence an issue?
- These methods can often be started at stationarity, so no burn-in
- If the underlying probability model is complex, simulating data will often not lead to acceptance. Thus need update for parts of the probability model (data augmentation)
- There are versions with varying $\epsilon$; see Bortot et al (JASA, 2007) for example
- There are now many hybrid versions of these approaches (e.g. ABC-within-Gibbs)
Variations on a theme – 4

Bortot et al (JASA, 2007) have a nice way to include $\epsilon$ in the ABC-MCMC approach

1. Now at ($\theta, \epsilon$)
2. Propose a move to ($\theta', \epsilon'$) according to $q((\theta, \epsilon) \rightarrow (\theta', \epsilon'))$
3. Generate $D'$ from model with parameters $\theta'$
4. Calculate
   \[
   h = \min \left( 1, \frac{\pi(\theta')\pi(\epsilon')q((\theta', \epsilon') \rightarrow (\theta, \epsilon))}{\pi(\theta)\pi(\epsilon)q((\theta, \epsilon) \rightarrow (\theta', \epsilon'))} \, \mathbb{1}(\rho(S', S) \leq \epsilon) \right)
   \]
5. Accept ($\theta', \epsilon'$) with probability $h$, else return ($\theta, \epsilon$)

Variations on a theme – 5

- The idea is to run the chain with typical values of $\epsilon$ being small
- Filter the series $\{(\theta_i, \epsilon_i)\}$ by restricting to $\{i : \epsilon_i < \epsilon_T\}$ after accepting values from the chain
  These values provide an estimate of $f(\theta|D)$ from values of $f(\theta|\rho \leq \epsilon_T)$ with weights given by $\pi(\epsilon)$
- The authors use priors of the form $\epsilon \sim \text{Exp}(\tau)$
SMC – 1

Repetitive sampling from the prior does not seem sensible, and various approaches have been designed to deal with this. Here is a version from Beaumont et al (Biometrika, 2009)

The aim is to

• perform weighted resampling of the points already drawn
• shrink $\epsilon$ as you go

In what follows,

• $D_0$ are the observed data
• $\epsilon_1 > \epsilon_2 > \cdots > \epsilon_T$ are given

SMC – 2

1. For iteration $t = 1$,
   1. For $i = 1, 2, \ldots, N$
      
      Simulate $\theta_i^{(1)} \sim \pi(\theta)$ and $D$ from the model with parameter $\theta_i^{(1)}$ until $\rho(D, D_0) < \epsilon_1$
      
      Set $w_i^{(1)} = 1/N$
      
      Calculate $\tau_2^2 = $ twice the empirical variance of the $\{\theta_i^{(1)}\}$
SMC – 3

2 For iteration $t = 2, \ldots, T$,
   
   For $i = 1, 2, \ldots, N$, repeat

   Choose $\theta_i^*$ from the $\theta_j^{(t-1)}$'s with probability $w_j^{(t-1)}$

   Generate $\theta_i^{(t)} \sim N(\theta_i^*, \tau^2_t)$, and $D$ from the model with parameter $\theta_i^{(t)}$

   until $\rho(D, D_0) < \epsilon_t$

   Set

   $$w_i^{(t)} \propto \frac{\pi(\theta_i^{(t)})}{\sum_{j=1}^N w_j^{(t-1)} \phi(\tau^{-1}_{t-1}(\theta_i^{(t)} - \theta_j^{(t-1)}))}$$

   Calculate $\tau^2_{t+1} = \text{twice the empirical variance of the } \{\theta_i^{(t)}\}$

SMC – 4

In the previous slide, $N(\mu, \sigma^2)$ denotes the Normal distribution with mean $\mu$ and variance $\sigma^2$, and $\phi(x)$ is the standard Normal density function.

Notes:
- Equally well, we can replace $D$ and $D_0$ with $S(D)$ and $S(D_0)$
- Can also replace the step

  $$\theta_i^{(t)} \sim N(\theta_i^*, \tau^2_t)$$

  with

  $$\theta_i^{(t)} \sim K(\cdot | \theta_i^*, \tau^2_t)$$

  where $K$ need not be a Gaussian kernel, but could be a $t$ distribution for example
SMC – 5: Rationale

- In step [1] we simulate from the prior and keep the $N$ closest points (so this is rejection-ABC). This set of points is drawn (roughly) from the posterior.
- Next, fit a density with kernel $K$

$$q(\theta) := \sum_{j=1}^{N} w_j^{(t-1)} K(\theta | \theta_j^{(t-1)}, \tau^2_\theta)$$

around the points, and resample values from this.
- then reduce the tolerance, $\epsilon$
- simulate data and their summary statistics
- weight each point by $\pi(\cdot)/q(\cdot)$ to allow for the fact that the points are not samples from the prior.