What we will do today

- The first ABC method
- Connection with sufficiency
- Examples
- Regression-based methods

29

The first ABC method

In the earlier methods we have assumed that we can hit the target. What happens if the acceptance probability is very small?
This gives us the first ABC method:

1. Generate $\theta \sim \pi(\cdot)$
2. Generate $\mathcal{D}'$ from the model with parameter θ
3. Accept θ if $\rho(\mathcal{D}', \mathcal{D}) < \epsilon$, where
   - ρ is a metric on the space of $\mathcal{D}$s
   - $\epsilon \geq 0$ is a parameter to be chosen

   Return to [1.]

30

Notes

- We can choose ρ to compare the data sets in a useful way
- If ρ is a metric, then $\rho(\mathcal{D}', \mathcal{D}) = 0 \iff \mathcal{D} = \mathcal{D}'$.
  Hence
  - $\epsilon = 0$ gives the exact answer
  - $\epsilon \to \infty$ reproduces the prior
  - $0 < \epsilon < \infty$ gives trade-off between accuracy and computability
- This method works for continuous data – Weiss and von Haeseler (1998) treated the frequentist case
- For the population genetics example we could use neighbourhoods of $\{S_n = s\}$ as the region.

31

ABC

If we were to use the full sequence data in that example, the target is very hard to hit. This suggests summarising the data.

This gives us the following ABC method (Pritchard et al, 1999):

1. Generate $\theta \sim \pi(\cdot)$
2. Generate $\mathcal{D}'$ from the model with parameter θ
3. Choose a set of summary statistics S of the data
   - Compute $S \equiv S(\mathcal{D})$, and $S' = S(\mathcal{D}')$
   - Accept θ if $\rho(S', S) < \epsilon$, where
     - ρ is a metric on the space of Ss

   Return to [1.]

32

The connection with sufficiency

Note that

$$\rho(S(\mathcal{D}),\, S(\mathcal{D}')) = 0 \;\not\Rightarrow\; \mathcal{D} = \mathcal{D}'$$

We are employing a dimension-reduction strategy here. There is a connection with sufficiency.

Recall that a statistic $S = S(\mathcal{D})$ is *sufficient* for the parameter θ if

$$\mathbb{P}(\mathcal{D}|S, \theta) \text{ is independent of } \theta$$

33

If S is sufficient for θ, then

$$\begin{aligned} f(\theta|\mathcal{D}) &\propto \mathbb{P}(\mathcal{D}|\theta)\,\pi(\theta) \\ &= \mathbb{P}(\mathcal{D}, S(\mathcal{D})|\theta)\pi(\theta) \\ &= \mathbb{P}(\mathcal{D}|S(\mathcal{D}), \theta)\mathbb{P}(S(\mathcal{D})|\theta)\pi(\theta) \\ &\propto \mathbb{P}(S|\theta)\pi(\theta) \\ &= f(\theta|S) \end{aligned}$$

We are really after a notion of *approximate sufficiency*

Research Question: If we had a measure of how close S is to sufficient, we should be able to metrise the difference between $f(\theta|\mathcal{D})$ and $f(\theta|S)$.

34

A Normal example – 1

Suppose $X_1, X_2, \ldots, X_n$ are iid N(μ, σ^2), with σ^2 known.

$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$ is sufficient for μ

Think of the prior as U(a, b), with $a \to -\infty, b \to \infty$

The posterior is truncated N$(\bar{X}, \sigma^2/n)$, restricted to (a, b)

For the ABC method, assume $\bar{X} = 0$. Then

1. Generate $\mu \sim \mathrm{U}(a, b)$
2. Generate $X_1, \ldots, X_n \sim \mathrm{N}(\mu, \sigma^2)$
3. Accept μ if $\rho(\bar{X}, \bar{X}_0 = 0) = |\bar{X}| \leq \epsilon$

A Normal example – 2

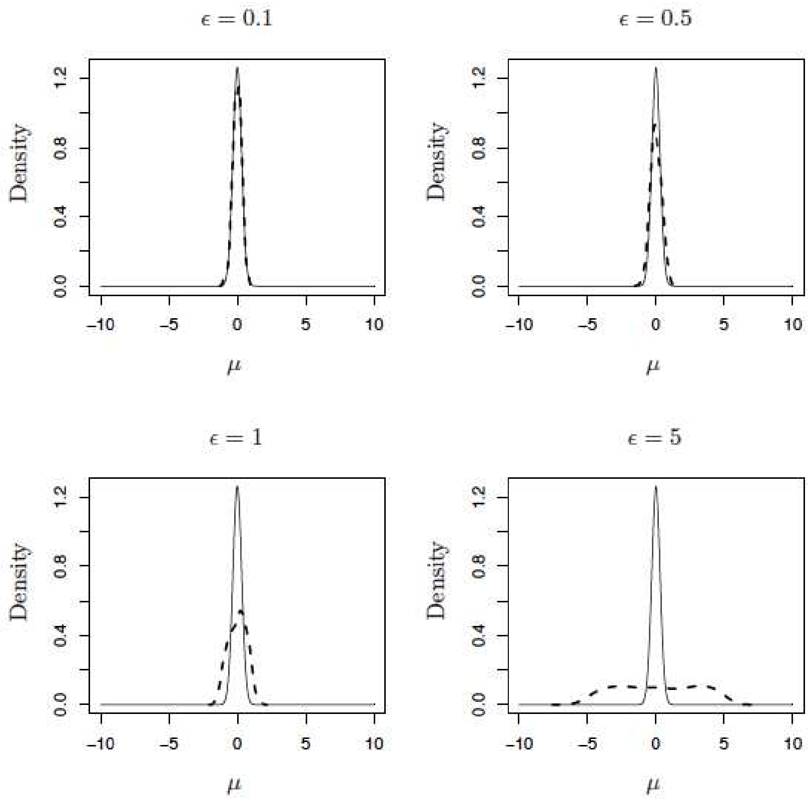

Figure 3.1: *Plots of the true posterior distribution for μ (solid line), and the ABC estimate using 1000 samples (dashed line). A value of $\sigma^2 = 1$ was used for all four plots.*

A Normal example – 3

The ABC density is proportional to

$$\mathbb{1}(a < \mu < b) \int_{-\epsilon}^{\epsilon} (2\pi\sigma^2/n)^{-1/2} \exp(-n(y-\mu)^2/2\sigma^2) dy$$

and the normalizing constant is $\int_a^b () d\mu$. We look at the case where $a \to -\infty, b \to \infty$. Then the normalising constant is 2ϵ, and the density is

$$f_\epsilon(\mu) = \frac{1}{2\epsilon} \left(\Phi\left(\frac{\epsilon - \mu}{\sigma/\sqrt{n}} \right) - \Phi\left(\frac{-\epsilon - \mu}{\sigma/\sqrt{n}} \right) \right)$$

Furthermore,

$$\mathbb{E}(\mu \,|\, |\bar{X}| \leq \epsilon) = 0, \quad \mathrm{Var}(\mu \,|\, |\bar{X}| \leq \epsilon) = \frac{\sigma^2}{n} + \frac{\epsilon^2}{3}$$

37

A Normal example – 4

Note that the variance shows *overdispersion*: the variance of the ABC method is larger than the true variance

Next we use a Taylor expansion to show that

$$d_{\mathrm{TV}}(f_\epsilon, f) := \frac{1}{2} \int_{-\infty}^{\infty} |f_\epsilon(\mu) - f(\mu)| d\mu \approx \frac{cn\epsilon^2}{\sigma^2} + o(\epsilon^2), \epsilon \to 0$$

Here, $c = \sqrt{\frac{2}{\pi}} e^{-1/2} \approx 0.4839$

Note: This example is from Richard Wilkinson's (2007) DAMTP PhD thesis

38

Motivation – 1

The idea is to replace the hard cut-off in the ABC with a soft version that exploits 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 ρ. Beaumont et al (2002) added

- smooth weighting
- regression adjustment

The method is insensitive to the value of ϵ, and allows m to be large.

Motivation – 2

We start from M observations (θ_i, S_i), where each θ_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 standardise 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 (θ_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 $||S - s_0||$ are the ones to use.

This can be improved by

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

Imagine that we have

$$\theta_i = \alpha + (s_i - s_0)^T \beta + \epsilon_i, i = 1, 2, \ldots, M \tag{3}$$

where the ϵ_i are uncorrelated $(0, \sigma^2)$ rvs

42

Motivation – 4

When $s_i = s_0$, θ_i is drawn from a distribution with mean

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

The least-squares estimator of (α, β) 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

43

Motivation – 5

$$X = \begin{pmatrix} 1 & s_{11} - s_{01} & \dots & s_{1m} - s_{0m} \\ \vdots & \vdots & & \vdots \\ 1 & s_{M1} - s_{01} & \dots & s_{Mm} - s_{0m} \end{pmatrix}$$

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 $\mathbb{E}(\theta|s_0) = \hat{\alpha} = M^{-1} \sum \theta_i^*$

44

Regression method – 1

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

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

One choice is the Epanechnikov kernel

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

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

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

45

Regression method – 2

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

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

- For local-constant regression, we set $\beta = 0$
- Then if $K_\epsilon(\cdot)$ is replaced by

  $$I_\epsilon(t) = \epsilon^{-1}\, \mathbb{1}(t \leq \epsilon)$$

  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

46

Choice of ϵ

- Set ϵ to be a quantile, P_ϵ, of the empirical distribution of simulated values of $||s_i - s_0||$
- The choice of ϵ 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

47