Selection of summary statistics for approximate Bayesian computation

David Balding and Matt Nunes

Institute of Genetics
University College London

COMPSTAT Paris, 26 August 2010

Research funded by UK EPSRC "Mathematics underpinning the life sciences" initative.

Approximate Bayesian Computation (ABC)

Assumptions:

- Dataset X is sampled from a known statistical model $M(\phi)$.
- Prior distribution for ϕ has density π.
- We can efficiently simulate from π and from $M(\phi)$.

Goal:

Approximate the posterior $\Pi(\phi|X)$ without computing the likelihood.

"Exact" Solution (Algorithm 0):

1. Simulate i.i.d. $\phi_i \sim \pi$, for $i = 1, \ldots, N$.
2. For each i, simulate a dataset $X_i \sim M(\phi_i)$.
3. If $X_i = X$, retain ϕ_i, otherwise discard.

The retained values form a random sample from $\Pi(\phi|X)$. By making N sufficiently large, any property of Π can be approximated to any desired accuracy.

A more realistic ABC

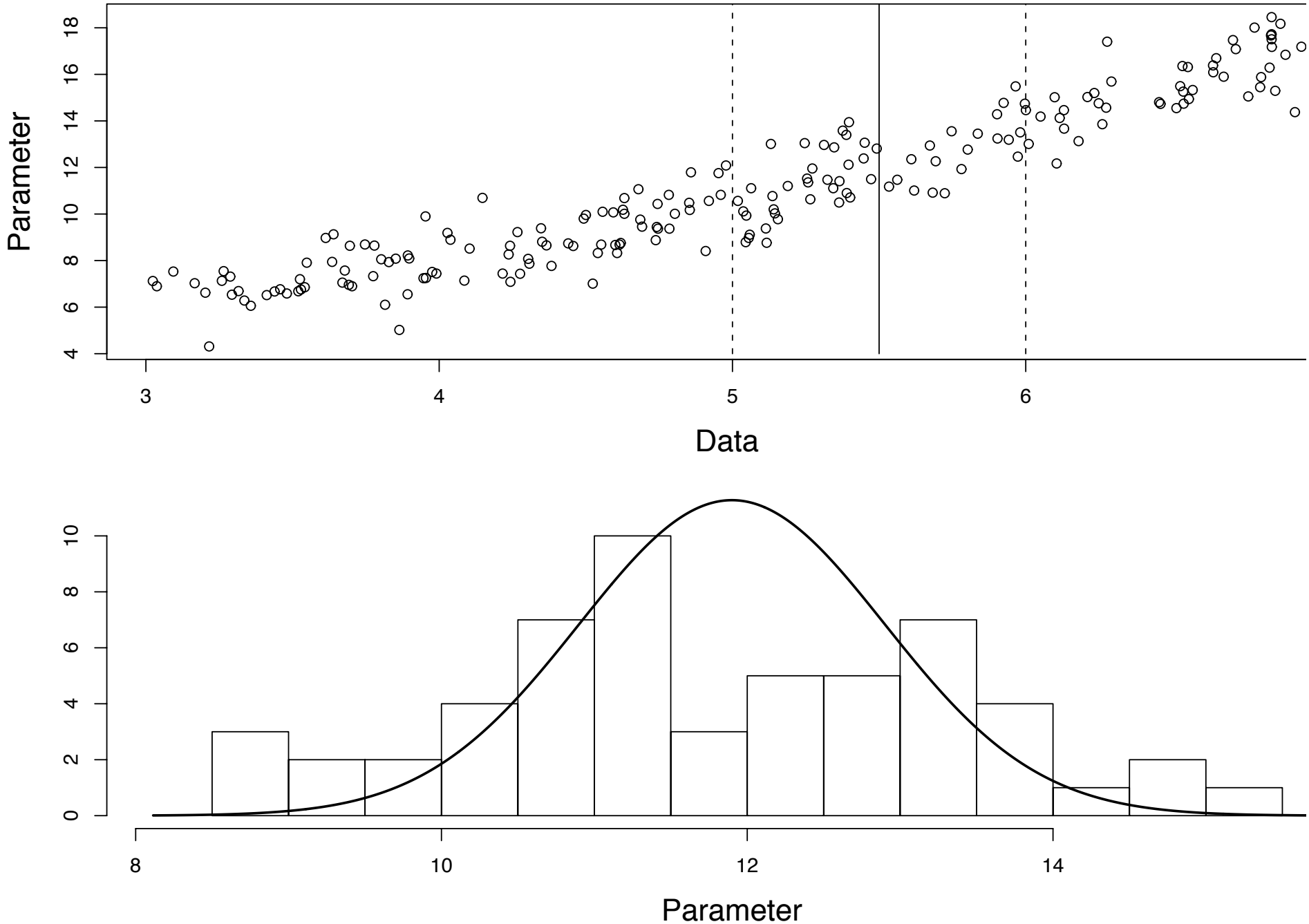
Equality of X_i with X is usually rare, so Algorithm 0 is impractical. Instead we require only that X_i and X are "close".

We define close in terms of a vector of summary statistics S such that datasets X_i and X convey approximately the same information about ϕ if and only if $S(X_i) \approx S(X)$.

Rejection-ABC Algorithm: Modify Algorithm 0 so that we retain ϕ_i whenever $||S(X_i) - S(X)|| < \delta$, where

- $||\cdot||$ denotes Euclidean distance;
- for fixed computing effort, δ specifies a bias-variance trade-off;
- can fix δ post-hoc to obtain a desired acceptance rate, say 1%.

The retained values form a random sample from $\Pi(\phi|S(X))$ which approximates the target posterior $\Pi(\phi|X)$.


Parameter
4
6
8
10
12
14
16
18
3
4
5
6
Data
0
2
4
6
8
10
8
10
12
14
Parameter

ABC: pros and cons

The Achilles' Heel of ABC is that it is difficult to quantify the error induced by only requiring $||S_i - S|| < \delta$ rather than $X_i = X$

- but can compute integrated square error (ISE) for similar simulated datasets.

ABC: pros and cons

The Achilles' Heel of ABC is that it is difficult to quantify the error induced by only requiring $||S_i - S|| < \delta$ rather than $X_i = X$

- but can compute integrated square error (ISE) for similar simulated datasets.

Because of this, ABC should only be used as a last resort.

- But it is the only available option in many settings involving complex models with many latent variables, and high-dimensional datasets.

ABC: pros and cons

The Achilles' Heel of ABC is that it is difficult to quantify the error induced by only requiring $||S_i - S|| < \delta$ rather than $X_i = X$

- but can compute integrated square error (ISE) for similar simulated datasets.

Because of this, ABC should only be used as a last resort.

- But it is the only available option in many settings involving complex models with many latent variables, and high-dimensional datasets.

Several "adaptive" variants have been proposed instead of rejection=ABC, the most important are

- Markov Chain Monte-Carlo ABC (Marjoram *et al.* 2003)
- Sequential Monte-Carlo ABC (Sisson *et al.* 2007; Beaumont *et al.* 2009).

Problem: how to choose summary statistics?

In practice they are selected by investigators on the basis of intuition and established practice in the field.

- To date, ABC has largely been applied in population genetics, where there are many well-established statistics that investigators know and love.
- ABC is related to "indirect inference" that evolved in econometrics (Gourieroux *et al.* 1993), and also to the generalised method of moments. These are focussed on classical estimators rather than Bayesian inference.
- ABC is now becoming used in infectious disease epidemiology and in systems biology.

Ideally S should be *sufficient* for ϕ. In practice this is unachievable, but (informally) we try to get as close to sufficiency as possible.

Approximate sufficiency

Joyce & Marjoram (2008) propose a notion of *approximate sufficiency* (AS) and use it to develop an algorithm for selecting good sets of summary statistics from a pool Ω.

- if S is sufficient then $\Pi(\phi|S) = \Pi(\phi|S \cup \{T\})$;
- so, given a current $S \subset \Omega$ they choose a random $T \in \Omega \backslash S$ and replace S with $S' = S \cup \{T\}$ if the ratio of the resulting posterior density estimates departs significantly from 1;
- if T is accepted, test all leave-one-out subsets of S';
- start with $S = \emptyset$; continue until no further change is possible.

First principled approach to selecting summary statistics for ABC.

Drawbacks include: no global optimum (dependence on random selection of new statistics to try); dependence on the threshold for a "significant" difference in ratio of posterior densities.

Partial Least Squares (PLS)

- Wegmann *et al.* (2009) suggest using PLS to derive orthogonal linear combinations of statistics from Ω that are maximally correlated with ϕ.
- For computational reasons, apply PLS to a 1% random subsample of the (ϕ_i, X_i) simulations.
- They propose leave-one-out cross validation to choose the optimal number of components.

Drawbacks of this approach include

- lack of interpretability of the PLS components;
- all PLS components used are given equal weight;
- PLS is applied to entire data space, whereas our interest is local to the observed X.

Minimum Entropy (ME)

We use a sample-based approximation to entropy as a heuristic for selecting summary statistics. Why entropy?

- widely-used measure of information
- related to variance, handles multi-modal distributions better
- not a property of direct interest.

We use *kth nearest neighbor entropy estimator* (Singh *et al.* 2003)

$$\log\left[\frac{\pi^{p/2}}{\Gamma(p/2+1)}\right] - \psi(k) + \log n + \frac{p}{n}\sum_{i=1}^{n}\log R_{i,k}$$

where p denotes the dimension of ϕ and $R_{i,k}$ is the Euclidean distance from ϕ_i to its kth nearest neighbour in the posterior sample, while $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function.

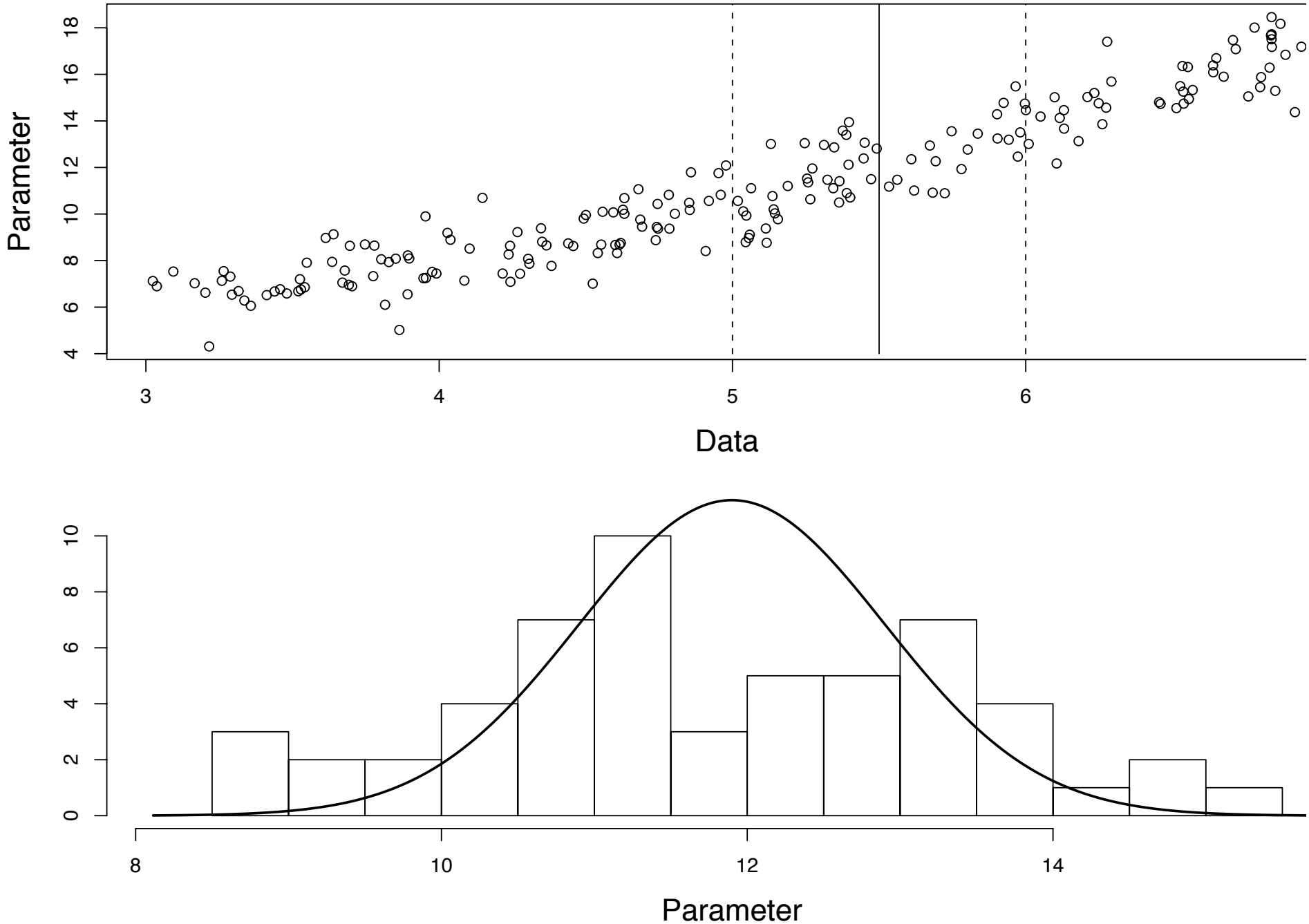
- Definition applies to multivariate case.
- We take $k = 4$ as suggested by Singh *et al.* (2003).
- For $|\Omega| < 10$ (say) can exhaustively consider all $S \subset \Omega$;
  - for larger Ω may need to limit e.g. to $S : |S| \leq k$.

Two-stage procedure

Stage 1: Find $S \subset \Omega$ that minimises the estimated entropy of $\Pi(\phi|S)$, using rejection-ABC.

Stage 2: Using S identified in Stage 1, take the k simulated datasets (below $k = 100$) that minimise $||S_i - S||$. For each $S \subset \Omega$, perform rejection-ABC and compute the RISE of $\Pi(\phi|S)$ for each of the k datasets, and hence select the subset of Ω that minimises MRISE.

Motivating intuition: our (large) set of (ϕ_i, X_i) simulations generates many X_i that are similar to the observed X, but for which the generating ϕ_i is known. Thus we can find the $S \subset \Omega$ that optimises the MRISE (or any preferred measure of accuracy of Π for ϕ) over these X_i. We have a "bootstrap" problem that we don't yet have a definition of "similar to", solved by using ME.


Parameter
4
6
8
10
12
14
16
18
3
4
5
6
Data
0
2
4
6
8
10
8
10
12
14
Parameter

Regression adjustment: mean

ABC algorithms can usually be improved by adjusting each accepted ϕ_i to correct for the (small) discrepancy between X_i and X (summarised by S_i and S, respectively). Beaumont *et al.* (2002) fitted a linear regression model and replaced ϕ_i with

$$\phi_i' = \hat{\alpha} + \hat{\epsilon}_i = \phi_i + (S_i - S)^T \hat{\beta},$$

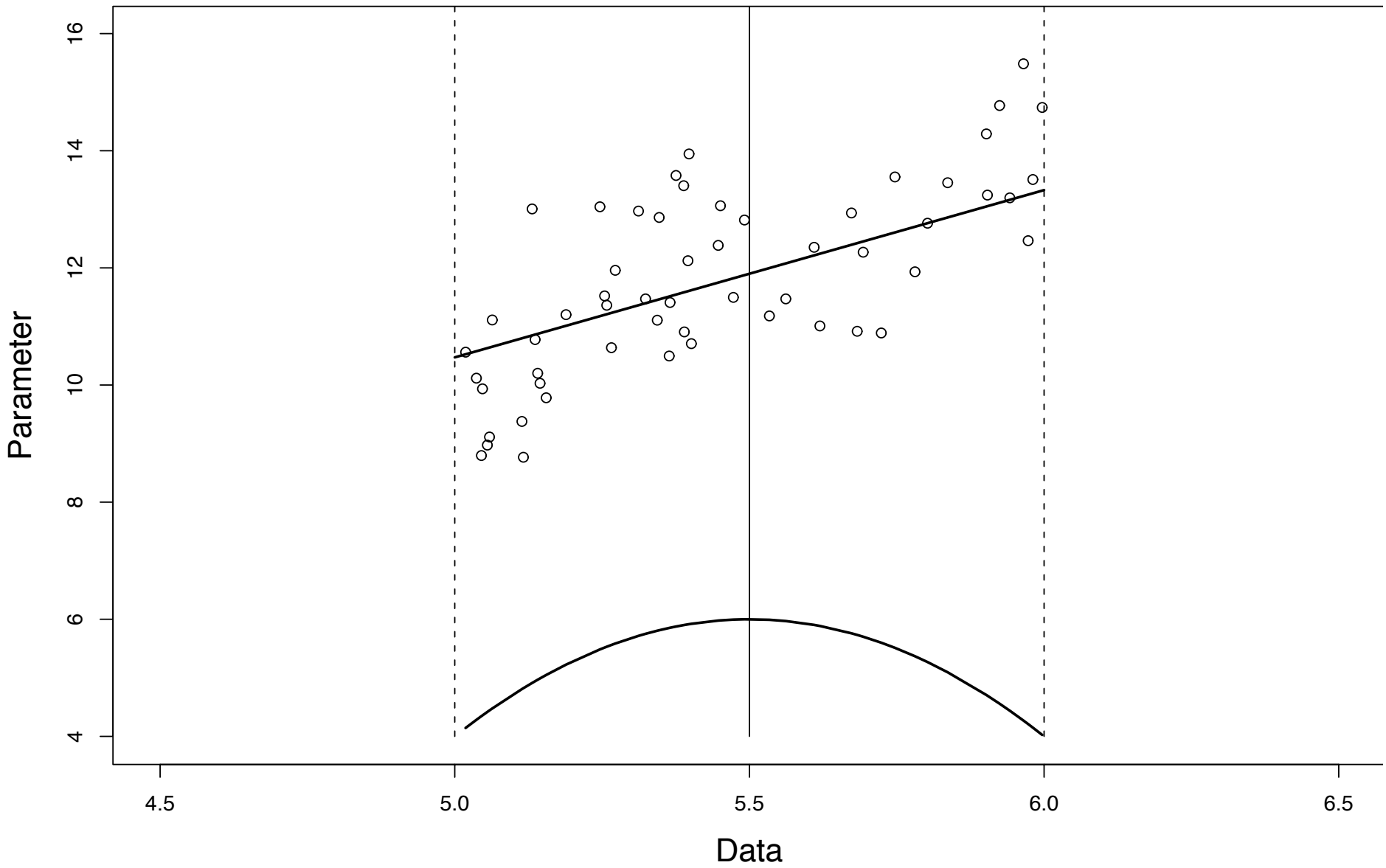
where $(\hat{\alpha}, \hat{\beta}) = (V^T W V)^{-1} V^T W \phi$ is the least squares estimator of the regression parameters and V is the design matrix of distances $(S - S_i)$. The weight matrix W is defined by

$$W_{ij} = \begin{cases} K(\|S_i - S\|), & i = j \\ 0 & \text{otherwise.} \end{cases}$$

with K the Epanechnikov kernel

$$K_\epsilon(t) = \begin{cases} 3(1-(t/\epsilon)^2)/4\epsilon, & t \leq \epsilon \\ 0 & t > \epsilon. \end{cases}$$

The W_{ij} are also applied to the ϕ_i' in approximating $\Pi(\phi|S)$.


Parameter
16
14
12
10
8
6
4
4.5
5.0
5.5
6.0
6.5
Data

Regression adjustment: variance

It may also be useful to adjust for systematic changes in the variance of ϕ_i' as S_i deviates from S, using a locally log-linear regression for the squared residuals from the mean adjustment:

$$\log(\hat{\epsilon_i}^2) = \alpha + (S_i - S)^T \beta + \varepsilon_i,$$

and estimate parameters with the same weight least-squares as above. Then we obtain the adjusted parameter values

$$\phi_i'' = \hat{\alpha} + \hat{\epsilon_i} \frac{\hat{\sigma}(S)}{\hat{\sigma}(S_i)}.$$

Feed-forward neural networks have also been proposed to make both mean and variance adjustments in the ABC setting (Blum & Francois, 2009)

Simulation study

- **Data:** 50 haplotypes (simplified coding of DNA sequences).
- **Model:** standard coalescent with infinite-sites mutation, implemented in *MS* software (Hudson 2002).
- **Targets of inference:** scaled mutation and recombination parameters, θ and ρ.
- **Prior:** $(\theta, \rho) \sim$ Unif(2,10) $\times$ Unif(0,10).
  - prior also used in simulations
  - unrealistic assumption but OK for method comparison.
- 10^6 datasets were simulated from which 100 were chosen at random to play the role of "observed" datasets.
- Rejection-ABC, with and without regression adjustments, was applied to each "observed" dataset for each S, accepting 10^4 $(= 1\%)$ of the simulated samples.

Simulation study: the pool of summary statistics, Ω

Statistic	Description
C_1	the number of segregating sites
C_2	a random draw from Unif[0,25]
C_3	the mean number of differences over all pairs of haplotypes
C_4	$25 \times$ (mean r^2 of pairs separated by $< 10\%$ of the simulated genomic region)
C_5	the number of distinct haplotypes
C_6	the frequency of the most common haplotype
C_7	the number of singleton haplotypes

- Each statistic was first standardised to unit variance.
- We compared all 127 non-empty subsets of the pool, to identify the subset that minimised Root Integrated Squared Error of the 10^4 points accepted by the ABC.
- We did this for θ and ρ separately, and then for joint estimation of (θ, ρ).

Performance of individual statistics

		C_1	C_2	C_3	C_4	C_5	C_6	C_7
θ	no adjustment	**1.75**	3.27	2.26	3.15	2.33	2.89	2.45
	mean	**1.75**	3.27	2.26	3.14	2.33	2.89	2.45
	mean+var.	**1.75**	3.27	2.26	3.14	2.33	2.89	2.45
ρ	no adjustment	3.93	3.95	3.93	3.92	**3.83**	3.84	3.88
	mean	3.92	3.95	3.93	3.92	**3.83**	3.84	3.89
	mean+var.	3.92	3.95	3.93	3.92	**3.83**	**3.83**	3.88
(θ, ρ)	no adjustment	**4.36**	5.19	4.62	5.10	4.55	4.89	4.65
	mean	**4.36**	5.19	4.62	5.10	4.56	4.89	4.65
	mean+var.	**4.36**	5.19	4.62	5.10	4.55	4.89	4.65

- RMISE (averaged over 100 "observed" datasets) of ABC inference using a single summary statistic.
- "noise" statistic C_2 is worst in each row; **bold** indicates best.
- No single statistic performs well for ρ.
- Regression adjustments of little use here.

Performance of methods of selecting sets of statistics

		best single	all six	PLS	AS	ME	two stage
θ	no adjustment	1.75	1.87	1.83	1.86	1.80	**1.70**
	mean	1.75	1.74	1.78	1.76	1.74	**1.68**
	mean+var.	1.75	1.70	1.75	1.76	1.70	**1.67**
ρ	no adjustment	3.83	3.59	3.91	3.68	3.54	**3.44**
	mean	3.83	3.33	3.37	3.83	3.56	**3.31**
	mean+var.	3.83	3.21	3.35	3.60	3.27	**3.17**
(θ, ρ)	no adjustment	4.36	4.81	4.15	-	4.03	**3.97**
	mean	4.36	4.83	4.08	-	4.06	**3.81**
	mean+var.	4.36	4.75	4.03	-	3.71	**3.66**

- Two-stage algorithm is best in each row.
- ME is 2nd best in 4 of 6 rows.
- $S =$ all six summary statistics (excluding C_2) performs well for univariate inference, not for bivariate.
- Both regression adjustments can help a lot.

Is the improvement of 2-stage algorithm significant?

		AS vs. 2-stage		PLS vs. 2-stage	
		ΔMRISE (s.e.)	p-value	ΔMRISE (s.e.)	p-value
θ	no adjust.	0.151 (0.024)	$< 10^{-8}$	0.130 (0.020)	$< 10^{-8}$
	mean	0.078 (0.029)	0.0042	0.099 (0.020)	$< 10^{-5}$
	mean+var.	0.097 (0.024)	$< 10^{-4}$	0.086 (0.023)	0.0002
ρ	no adjust.	0.239 (0.048)	$< 10^{-5}$	0.471 (0.070)	$< 10^{-9}$
	mean	0.525 (0.090)	$< 10^{-7}$	0.059 (0.031)	0.0300
	mean+var.	0.426 (0.081)	$< 10^{-6}$	0.181 (0.042)	$< 10^{-4}$
(θ, ρ)	no adjust.			0.185 (0.053)	0.0004
	mean			0.270 (0.065)	$< 10^{-4}$
	mean+var.			0.369 (0.071)	$< 10^{-6}$

Difference in MRISE (ΔMRISE) for the pair of methods indicated in the column heading, together with its standard error and p-value (one-sided t_{99} test).

Most frequently selected subsets (unadjusted inference)

	true RISE	#	est. entropy	#	MRISE	#
θ	$\{C_1, C_3\}$	26	$\{C_1, C_3\}$	31	$\{C_1, C_4\}$	60
	$\{C_1, C_4\}$	23	$\{C_1, C_4\}$	29	$\{C_1, C_3\}$	34
	$\{C_1, C_5\}$	12	$\{C_1, C_3, C_4\}$	17	$\{C_1, C_4, C_5\}$	33
ρ	$\{C_1, C_4, C_5\}$	34	$\{C_1, C_4, C_5\}$	33	$\{C_1, C_4, C_5\}$	73
	$\{C_1, C_5\}$	17	$\{C_1, C_5\}$	23	$\{C_1, C_3, C_4, C_5\}$	40
	$\{C_1, C_3, C_4, C_5\}$	15	$\{C_3, C_4, C_5\}$	16	$\{C_1, C_5\}$	23
(θ, ρ)	$\{C_1, C_4, C_5\}$	29	$\{C_1, C_5\}$	41	$\{C_1, C_4, C_5\}$	76
	$\{C_1, C_3, C_5\}$	17	$\{C_1, C_4, C_5\}$	40	$\{C_1, C_3, C_4, C_5\}$	44
	$\{C_1, C_5\}$	16	$\{C_1, C_3, C_5\}$	27	$\{C_1, C_5\}$	39

Top three S in frequency (out of 100 observed datasets) for which ABC inference achieved within 1% of the optimal value of: true RISE, estimated entropy, and MRISE (over 100 nearest datasets).

- Different S are optimal for different datasets.
- Justifies search for optimal statistics locally to observed X.
- Suggests averaging posterior approximations over several S.

Further work

- Restricting attention to subsets of Ω, each statistic having equal weight, is computationally convenient but restrictive.
- Exhaustive search over all $S \subseteq \Omega$ is expensive (10 hours per dataset versus 6 min, 3 min and 2 min for ME, AS and PLS).
  - Could restrict attention to S that are *a priori* plausible, and/or that perform well under ME.
  - Could implement an iterative scheme such as AS method but with superior MRISE criterion for accepting updates.
  - Could generalise to continuous updates of a weight vector.
- Can consider orthogonalising statistics.
- All the above is applied only to rejection-ABC; could also be applied to MCMC-ABC or SMC-ABC, with learning of the weight vector in addition to existing updates.

I apologise that there is no paper in the proceedings volume; software is available on request and a manuscript will shortly appear in *Statistics Applied in Genetics and Molecular Biology*.

And finally some light entertainment ...

AR TEMPLETON, Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation, Molecular Ecology **18**, *319–331 (2009).*

- "NCPA can yield strong inference whereas ABC yields weak inference and there is no way within ABC of identifying or correcting for this source of false-positives"
- "... disadvantage of ABC is that the interpretative set cannot be validated"
- "The field of statistics focuses upon the error induced by sampling from a population of inference"
- "In ABC, S_i is the long-term statistic, and ... evolutionary stochasticity and ... sampling error is taken into account via the computer simulation. However, S is the current generation statistic, and it is treated as a fixed constant ... thereby violating the known sources of error"

"As Fig. 3 illustrates, the ABC method cannot distinguish between a truly good fitting model, an uninformative model, or an over-determined model. This is the danger of using only local relative probabilities rather than true posterior distributions."

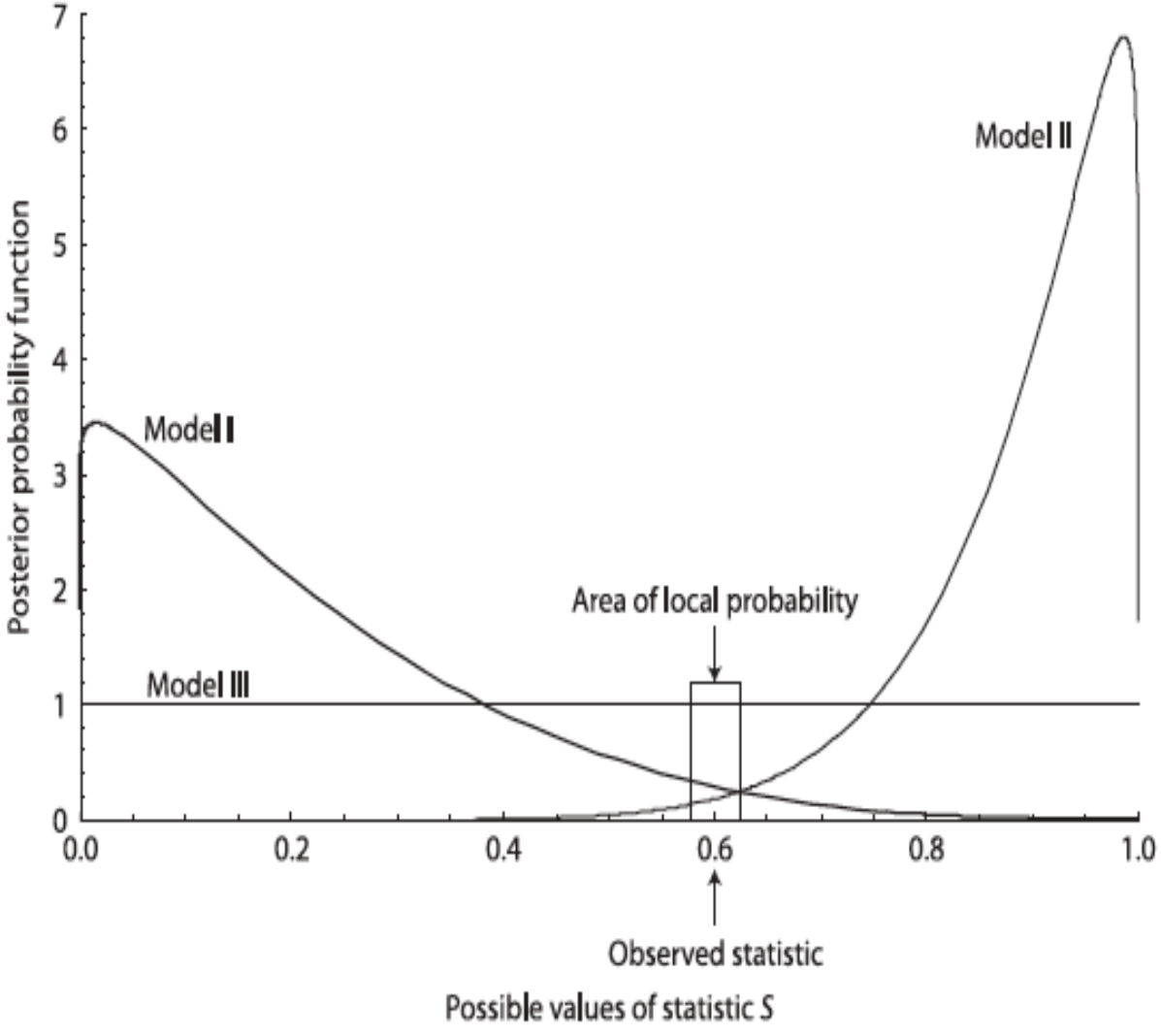


Fig. 3 Hypothetical posterior distributions for three models for a univariate statistic and an area around the observed value of statistic where local probabilities are evaluated.

- "The impact of treating S as a fixed constant is to increase statistical power as an artefact."
- "Ignoring the sampling error of S undermines the statistical validity of all inferences made by the ABC method."
- "Hence, the 'posterior probabilities' that emerge from ABC are not co-measureable. This means that it is mathematically impossible for them to be probabilities."
- "Thus, the final product of the ABC analysis are numbers that are devoid of statistical meaning. The ABC method is not capable of even weak statistical inference."

Many concerned citizens got together to try to put right as much as we could of Templeton's nonsense:

> *Beaumont MA, Nielsen R, Robert C, Hey J, Gaggiotti O, Knowles L, Estoup A, Panchal M, Corander J, Hickerson M, Sisson SA, Fagundes N, Chikhi L, Beerli P, Vitalis R, Cornuet JM, Huelsenbeck J, Foll M, Yang ZH, Rousset F, Balding D, Excoffier L (2010). In defence of model-based inference in phylogeography MOL ECOL 19(3), 436-446 doi:10.1111/j.1365-294X.2009.04515.x.*

only to discover to our dismay that he published the same rubbish again in the March 2010 issue of PNAS, without acknowledging our rebuttal (appeared Jan 2010).

It is difficult to defend science against pseudo-science when we must admit that an influential person can publish garbage in a top journal, and be recognised as a leading scientist with so little understanding of basic ideas of inference.