Non-linear Regression Approaches in ABC

Michael G.B. Blum

Lab. TIMC-IMAG, Université Joseph Fourier, CNRS, Grenoble, France

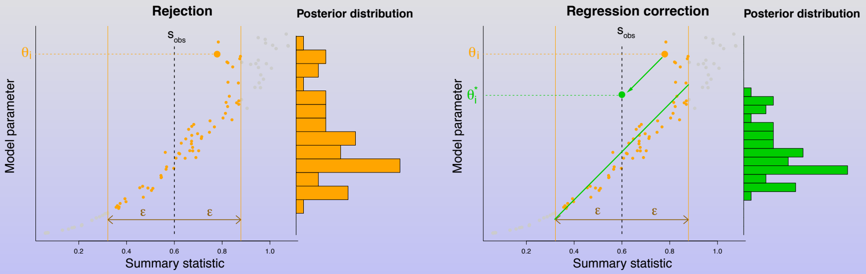
ABCiL, May 5 2011

Overview

1. Why using (possibly non-linear) regression adjustment in ABC : theoretical arguments
2. Methods for non-linear regression adjustment
3. Examples

Correction adjustment

Beaumont et al. Genetics 2002



Adapted from Csilléry et al. TREE 2010

If you prefer the math

Beaumont et al. Genetics 2002

- A model of local regression

$$\theta_i | \mathbf{s}_i = m(\mathbf{s}_i) + \epsilon_i$$

- Local linear approximation

$$m(\mathbf{s}_i) = \alpha + \mathbf{s}_i^t \beta$$

- Adjustment

$$\theta_i^* = \hat{m}(\mathbf{s}_{obs}) + \tilde{\epsilon}_i,$$

where $\tilde{\epsilon}_i$ are the empirical residuals.

Main theorem

Blum, JASA 2010

Asymptotic bias of the estimates of the posterior $\hat{g}_j(\theta|\mathbf{s}_{obs})$, $j = 0$ (rejection),1 (linear adj.), 2 (quadratic adj.)

$$C_j\,\varepsilon^2$$

Asymptotic variance of $\hat{g}_j(\theta|\mathbf{s}_{obs})$

$$\frac{C'}{np(\mathbf{s}_{obs})\varepsilon^d}$$

where d is the dimension of the summary statistics and n is the number of simulations.

Remark 1 : The curse of dimensionality

$$\text{Minimals MSE} = O(n^{-4/(d+5)}).$$

The rate at which the minimal MSEs converges to 0 decreases importantly (at least theoretically) as the dimension d of $\mathbf{s}_{obs}$ increases.

Possible solution

- Projecting the summary statistics on a lower dimensional subspace

Remark 2 : Comparison between the estimators with and without adjustment

When the model

$$\theta_i = m(\mathbf{s}_i) + \epsilon_i$$

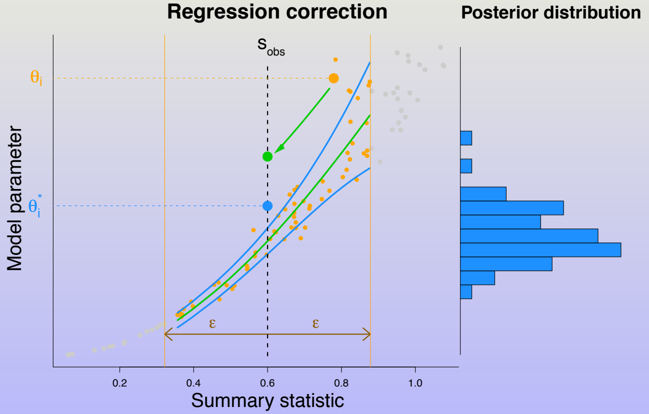
is homoscedastic in the vicinity of $\mathbf{s}_{obs}$, then
bias (quadratic adj.)$\leq$ bias (linear adj.)$\leq$ bias (without adj.)

Solutions

- Makes the model more homoscedastic : transformations of sum stats and parameters (not pursued here, see Blum JASA 2010)
- Provides a more flexible regression model : non-linear and heteroscedastic regression

Non-linear and heteroscedastic regression adjustment

Blum and François, Stat. Comput. 2010



Non-linear and heteroscedastic regression adjustment

Blum and François, Stat. Comput. 2010

- Innovation 1 : an heteroscedastic model of local regression

$$\theta_i|\mathbf{s}_i = m(\mathbf{s}_i) + \sigma(\mathbf{s}_i)\epsilon_i$$

- Innovation 2 : non linear function for m and σ
- Neural nets for m and σ for projecting on a lower dimensional subspace
- Heteroscedastic adjustment

$$\theta_i^* = \hat{m}(\mathbf{s}_{obs}) + \frac{\hat{\sigma}(\mathbf{s}_{obs})}{\hat{\sigma}(\mathbf{s}_i)}\tilde{\epsilon}_i,$$

Fitting neural networks

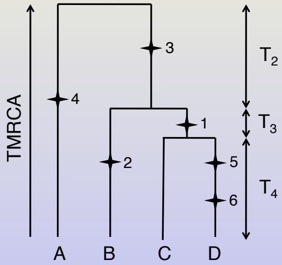
- Fit M (typically $M = 10$) neural networks and consider the median to obtain $\hat{m}$.
- Consider M regression model for fitting the conditional variance $\sigma(\cdot)$

$$\log((\theta_i - \hat{m}(\mathbf{s}_i))^2) = \log \sigma^2(\mathbf{s}_i) + \xi_i.$$

Example 1 : Coalescent model in population genetics

Nunes and Balding, SAGMB 2010

Segregating sites	Number of individuals
123456	
A 000100	1
B 011000	2
C 101000	6
D 101011	1



Model without recombination

- Inter-coalescence times $T_i \rightsquigarrow \text{Exp}(i(i-1)/2), i = 2, \ldots, n$
- Superimpose mutation using a Poisson process of rate $\theta/2$

Example 1 : summary statistics

$n = 10^6$, $n_{accepted} = 10^4$.

- C_1 Number of seg sites
- C_2 Unif. variable
- C_3 Mean number of differences over all pairs of haplotypes
- C_4 mean r^2
- C_5 Number of distinct haplotypes
- C_6 Frequency of the most common haplotype
- C_7 Number of singleton

Example 1 : Estimation of θ

$$\mathrm{RSSE} = \sqrt{\frac{1}{n_{accepted}} \sum_{\mathrm{Accepted\ points}} \|\theta_i - \theta\|_2^2}$$

$$\mathrm{MRSSE} = \mathrm{Average(RSSE)}$$

Relative MRSSE w.r.t. ABC with C_1 (number of seg. sites)

	Single sum stats								Selection of sum stat		Projection	
	C1	C2	C3	C4	C5	C6	C7	All 6	AS	2-stage	PLS	NN
No adj.	0	92	21	86	28	35	39	6	6	-3	5	
Homo. Linear adj	-	-	-	-	-	-	-	2	1	-4	2	
Hetero linear adj.	-	-	-	-	-	-	-	2	1	-5	0	1

PLS (Partial Least squares, Wegmann et al., Genetics 2009)

AS (Approximate Sufficiency, Joyce and Marjoram, SAGMB 2008)

2-stage (Entropy-based method, Nunes and Balding, SAGMB 2010)

Example 1 : Estimation of θ and ρ

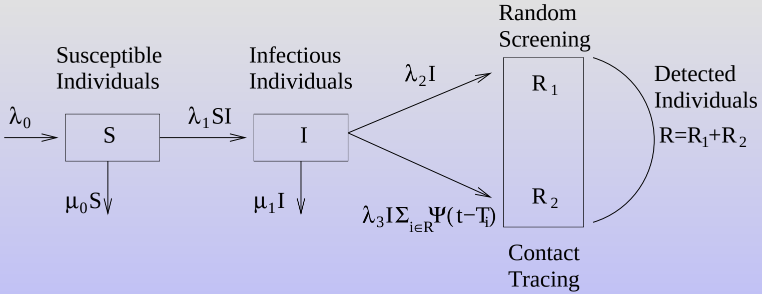
Relative MRSSE w.r.t. ABC with C_1 (number of seg. sites)

	Single sum stats								Selection of sum stat		Projection	
	C1	C2	C3	C4	C5	C6	C7	All 6	AS	2-stage	PLS	NN
No adj.	0	18	5	15	2	4	5	-7		-10	-5	
Homo. linear adj	-	-	-	-	-	-	-	-9		-14	-7	
Hetero. linear adj.	-	-	-	-	-	-	-	-15		-19	-8	-17

- Curse of dimensionality is not a severe issue here : ‘All 6’ performs good
- Homo. adjustment improves the results and hetero. adj. even further
- Projection with neural networks performs almost as good as the extremely time-consuming, but efficient, ‘2-stage’ method

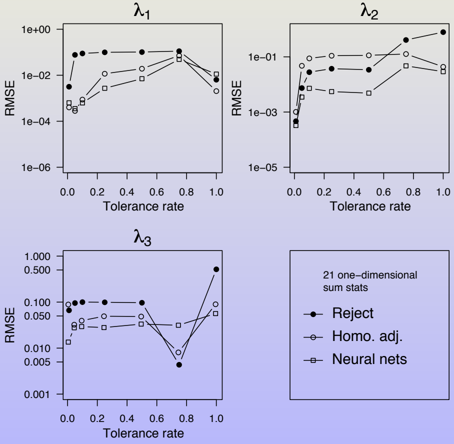
Example 2 : Compartmental model in epidemiology

Blum and Tran Biostatistics 2010



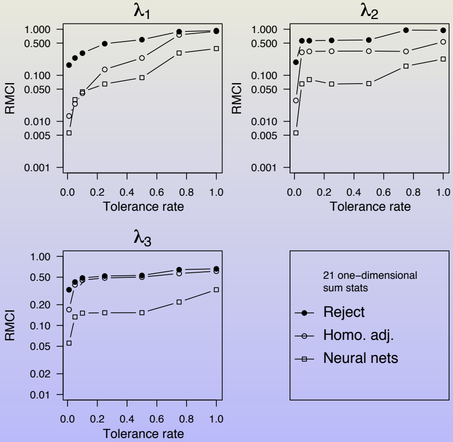
Example 2 : Mean square error of point estimates

Adjustments reduce RMSE (Rescaled Mean Squared Error)



Example 2 : Width of credibility intervals

Adjustments shrink the posterior



RMCI is Rescaled Mean Credibility Interval

Conclusions

- The curse of dimensionality might be a less severe problem than suggested by theoretical arguments

  Scott (1992), in the context of multivariate density estimation, argued that conclusions arising from the same kind of theoretical arguments were in fact much more pessimistic than the empirical evidence.
- Adjustments based on non linear heteroscedastic regression models shrink the posterior distribution
- Heteroscedastic regression models can be used with linear regression models (Nunes and Balding, SAGMB 2010)

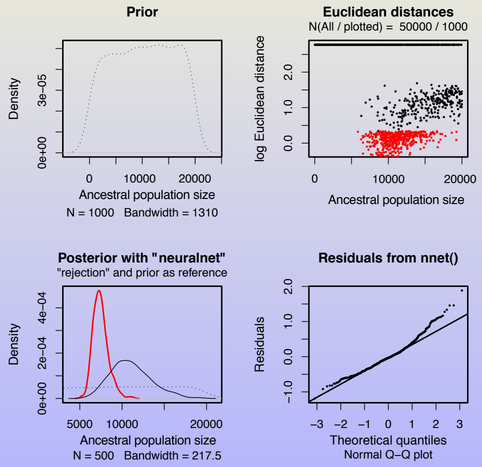
If you are not convinced....

You can use the R *abc* package to make your own opinion.
http://cran.r-project.org/web/packages/abc/index.html

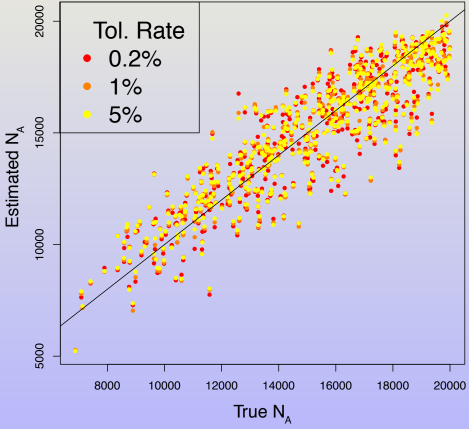
Implements various functions for parameter estimation, model selection as well as cross-validation tools.

Parameter inference with the R package

Effective population size in a coalescent model



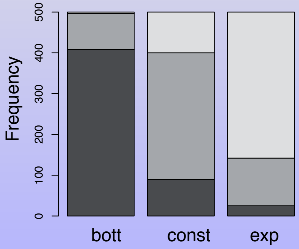
Cross validation for parameter inference



Cross validation for model selection

Confusion matrix: How many times the predicted models are the same as the true models?

	bott	const	exp
bott	408	89	3
const	90	310	100
exp	25	117	358



Collaborators

Katy Csilléry, Grenoble

Olivier François, Grenoble

Viet-Chi Tran, Lille