CONDITIONING ON PARAMETER POINT ESTIMATES IN APPROXIMATE BAYESIAN COMPUTATION

BY

Emilie Haon–Lasportes
Florence Carpentier
Olivier Martin
Etienne K. Klein
Samuel Soubeyrand

Research Report No. 45
July 2011

Unité de Biostatistique et Processus Spatiaux
Institut National de la Recherche Agronomique
Avignon, France
http://www.biosp.org

Conditioning on parameter point estimates in approximate Bayesian computation

E. Haon–Lasportes*, F. Carpentier†, O. Martin, E. K. Klein and S. Soubeyrand

INRA, UR546 Biostatistics and Spatial Processes, F-84914 Avignon, France

July 29, 2011

Abstract: The method that we propose allows to reduce the dimension of the summary statistics in approximate Bayesian Computation. It transforms the summary statistics into point estimates of the parameters which are then used as new summary statistics. On a toy example, we show that the dimension of the summary statistics and their handicaps are problems that are not managed correctly by classical ABC procedures. The proposed transformation leads to optimal results with regard to the posterior distribution, namely similar results to the sufficient statistics.

Keywords: Implicit stochastic model, unknown parameters, summary statistics, distance, point estimates

1 Introduction

Today, there exists a flourishing literature where mechanistic models are built to describe and infer complex dynamics; see for instance Ovaskainen and Hanski (2004); Real and Biek (2007); Soubeyrand et al. (2009) in population dynamics and Estoup et al. (2004) in population genetics. Among these mechanistic models some are stochastic implicit models: they can be used to generate data sets but the probability distribution of the state variables is untractable (Diggle and Gratton, 1984). Approximate Bayesian Computation (ABC) procedures have been especially developed to make Bayesian inference in the case of such implicit models (Beaumont et al., 2002, 2009; Blum and François, 2010; Marjoram et al., 2003; Sisson et al., 2007; Wegmann et al., 2009; Marin et al., 2011).

*Corresponding Author: Emilie.HaonLasportes@avignon.inra.fr

†Current address: AgroParisTech /INRA MIA UMR 518, Paris, France

The basic ABC procedure (ABC–rejection) is carried out as follows: (i) sets of parameters are independently simulated under prior distributions; (ii) for each set of parameters a data set with the same structure than the observed data set is simulated under the implicit model; (iii) the simulated parameters providing the simulated data sets closest to the observed data set are retained to build the posterior distribution. The closeness of the data sets is usually quantified with a distance between summary statistics computed from observed and simulated data. The use of summary statistics to compare the data sets in lower-dimension spaces (see e.g. Marjoram et al., 2003) may lead to difficulties in the implementation of ABC (ABC–rejection and other ABC procedures). This is the topic of the present article.

The presentation of the ABC–rejection procedure by Pritchard et al. (1999) in a population genetics context marks the emergence of ABC methods; even if Rubin (1984) previously proposed this procedure (without mentioning the use of statistics) but his proposal did not resound at that time. Since 1999, many improvements of the procedure concerning the exploration of the parameter space or the derivation of the posterior distributions were published. For instance, smoothing was used instead of acceptance/rejection (Beaumont et al., 2002; Blum and François, 2010), regression models were incorporated in the procedure (Beaumont et al., 2002; Blum and François, 2010), and iterative algorithms based on Markov chains or importance sampling were proposed instead of the independent sampling of the parameters (Marjoram et al., 2003; Beaumont et al., 2009; Wilkinson, 2008).

In the existing ABC procedures, another critical aspect for the accuracy of the inferences concerns the choice of the summary statistics and the choice of the distance between simulated and observed statistics. Because ABC procedures are applied to stochastic implicit model which are analytically intractable, determining theoretically the set of minimal sufficient statistics is cumbersome. In practice, when it is possible, a list of statistics expected to be strongly related with the unknown parameters is built based on the expertise of the analyst. But when a previous knowledge is not available, the choice of the summary statistics becomes a problem: some statistics may not be related to any parameter and are un-informative, some statistics may carry the same information for the parameters and are redundant. Several solutions have been already proposed to bear with some of these difficulties, i.e. to transform statistics or to reduce the dimension of the statistics, mainly trying to deal with the curse of dimensionality (the rate of convergence of the MSE criterion decreases as the dimension of the statistics increases). For instance, Joyce and Marjoram (2008) propose to select a subset of statistics based on relative ϵ-sufficiency; Wegmann et al. (2009) propose to transform the statistics into the axes of a PLS regression; Nunes and Balding (2010) propose to optimize the selection of summary statistics minimizing the average square error of the posterior distribution of the parameter; Fearnhead and Prangle (2010) propose a construction of summary statistics in a semi–automatic manner, optimal summary statistics being the posterior means of the parameters.

Despite all these recent propositions, the impact of nonlinearity, redundancy, uninformativeness, high dimensionality has been poorly investigated. In section 2 we propose

to partially fill this gap using ABC–rejection on a toy example. In Section 3 we propose a solution in the vein of the solution of Fearnhead and Prangle (2010): transforming the summary statistics into point estimates of the parameters by fitting an empirical model from the simulated data. Here, we use a regression model where the parameters are the response variables and the summary statistics are explanatory variables. The empirical model is known to be unsatisfactory from a scientific viewpoint as it is not mechanistic contrarily to the implicit model, but it is only used to get point estimates of the parameters which are then used as new summary statistics. In Section 4 we apply ABC–rejection and ABC–Beaumont in the same toy example incorporating our transformation and compare results.

2 Influence of the statistics in the ABC–rejection procedure

2.1 The ABC–rejection procedure

Consider observed data $\mathcal{D} \in \mathbb{D}$ which are assumed to be generated under the stochastic model $\mathcal{M}_\theta$ parametrized by $\theta \in \Theta$ with prior density π. The data space $\mathbb{D}$ and the parameter space Θ are both included in multidimensional sets of real vectors. The posterior distribution $p(\theta \mid \mathcal{D})$ can be estimated using the following ABC–rejection algorithm:

A1. Carry out the next two steps for i in $\{1, \ldots, I\}$, independently

- Generate θ_i from π and simulate $\mathcal{D}_i$ from $\mathcal{M}_{\theta_i}$.
- Accept θ_i if $\mathcal{D}_i = \mathcal{D}$, reject it otherwise.

The set of accepted θ_i form a sample from the posterior distribution

$$p(\theta \mid \mathcal{D}) = \frac{f(\mathcal{D} \mid \theta)\pi(\theta)}{\int_\Theta f(\mathcal{D} \mid \alpha)\pi(\alpha)d\alpha},$$

where $f(\mathcal{D} \mid \theta)$ is the conditional probability distribution function of $\mathcal{D}$ given θ, i.e. the (un-tracktable or unknown) likelihood corresponding to the model $\mathcal{M}_\theta$.

In practice, this algorithm can rarely be used: the probability of generating $\mathcal{D}_i$ equal to $\mathcal{D}$ is very low when the dimensionality of the data space $\mathbb{D}$ is large and this probability is even zero for continuous data. To circumvent this difficulty two ideas have been applied: the introduction of a tolerance threshold and the replacement of the raw data by statistics. This leads to the following ABC–rejection algorithm which will be extensively used in this article (Pritchard et al., 1999):

A2. Carry out the next three steps for i in $\{1, \ldots, I\}$, independently

- Generate θ_i from π and simulate $\mathcal{D}_i$ from $\mathcal{M}_{\theta_i}$.
- Compute the summary statistics $S_i = s(\mathcal{D}_i)$, where s is a function from $\mathbb{D}$ to the space $\mathbb{S}$ of statistics.

– Accept θ_i if $d(S_i, S) \leq \epsilon$, where d is a distance over $\mathbb{S}$ and ϵ is a tolerance threshold for the distance between the observed statistics $S = s(\mathcal{D})$ and the simulated ones S_i.

The set of accepted parameters, say $\Theta_{\epsilon,I} = \{\theta_i : d(S_i, S) \leq \epsilon, i = 1, \ldots, I\}$, forms a sample from the posterior distribution

$$p_\epsilon(\theta \mid S) = \frac{\left(\int_{B(S,\epsilon)} g(z \mid \theta) dz\right) \pi(\theta)}{\int_\Theta \left(\int_{B(S,\epsilon)} g(z \mid \alpha) dz\right) \pi(\alpha) d\alpha},$$

where $g(S \mid \theta)$ is the conditional probability distribution function of S given θ.

When ϵ tends to zero, $p_\epsilon(\theta \mid S)$ was proved to be a good approximation of the posterior distribution conditional on the statistics (Blum, 2010), i.e.

$$p(\theta \mid S) = \frac{g(S \mid \theta)\pi(\theta)}{\int_\Theta g(S \mid \alpha)\pi(\alpha) d\alpha},$$

and the sample $\Theta_{\epsilon,I}$ of accepted parameters is approximately distributed under this posterior distribution. In addition, if the sufficient statistics are used, $g(S \mid \theta) = f(\mathcal{D} \mid \theta)$ and $\Theta_{\epsilon,I}$ is approximately a sample from the classical posterior distribution $p(\theta \mid \mathcal{D})$ conditional on the data.

In practice, instead of the tolerance threshold ϵ, the analyst selects a tolerance rate $\tau \in]0, 1]$ and obtain ϵ as a function of τ: $\epsilon = \epsilon(\tau)$. A small value is generally chosen for τ. Typically, τ is 0.001 (e.g. 10^3 accepted parameters among $I = 10^6$ simulations).

In many applications (see e.g. Estoup et al., 2004; Hamilton et al., 2005; Guillemaud et al., 2009), the statistics are quantitative variables and the distance $d(S_i, S)$ is the Euclidean distance up to a standardization of the components of the statistics (i.e. each component of S_i and S is divided by the corresponding empirical standard deviation calculated by using the I simulations).

2.2 Comparison of the posterior distributions obtained with various statistics

We first investigate the effect of the choice of the summary statistics on the approximation of the posterior distribution. For this purpose we consider a toy example where the minimal sufficient statistics are known and we study the effects of nonlinearity, redundancy, increase in the dimension and lack of information in the summary statistics.

Toy example: The observed data set $\mathcal{D} = (X_1, \ldots, X_{40})$ is a sample of 40 independent normal variables with expectation $\mu \sim \mathcal{U}_{[-2,2]}$ and standard deviation $\sigma \sim \mathcal{U}_{]0,4]}$. Similarly, the prior for the simulated parameters, $\theta_i = (\mu_i, \sigma_i), i = 1 \ldots, I$, is the uniform distribution

over $[-2,2]\times]0,4]$ and each simulated data $\mathcal{D}_i$ is a sample of 40 independant normal variables with parameter θ_i. Six alternative functions s for computing the summary statistics are considered:

- Minimal sufficient statistics:

$$s_1(\mathcal{D}) = (\bar{X}_{1:40}, SD_{1:40}),$$

  where $\bar{X}_{i:j}$ and $SD_{i:j}$ are, respectively, the empirical mean and standard deviation of the (sub-)sample $(X_i, \ldots, X_j)$, $1 \leq i < j \leq 40$;

- Statistics with nonlinearity:

$$s_2(\mathcal{D}) = (\exp(\bar{X}_{1:40}), (SD_{1:40})^2);$$

- Statistics with redundancy:

$$s_3(\mathcal{D}) = (\bar{X}_{1:40}, \bar{X}_{1:20}, \bar{X}_{21:40}, SD_{1:40});$$

- Statistics with an increase in the dimension:

$$s_4(\mathcal{D}) = (\bar{X}_{1:10}, \bar{X}_{11:20}, \bar{X}_{21:30}, \bar{X}_{31:40}, SD_{1:40});$$

- Statistics with un-informativeness:

$$s_5(\mathcal{D}) = (\bar{X}_{1:40}, SD_{1:40}, B_1, B_2),$$

  where B_1 and B_2 are two independent random variables drawn from the beta distribution with shape parameters 0.1 and 0.1. Each time s_5 is applied to a new data set, new variables B_1 and B_2 are simulated independently. These random variables are added but bring no information;

- Statistics gathering all handicaps together:

$$s_6(\mathcal{D}) = (\bar{X}_{1:10}, \bar{X}_{11:20}, \bar{X}_{21:30}, \bar{X}_{31:40}, (SD_{1:40})^2, (SD_{1:20})^2, (SD_{21:40})^2, B_1, B_2).$$

We use the same simulations $\{(\theta_i, \mathcal{D}_i) : i = 1, \ldots, I\}$ for the six sets of summary statistics $s_1, \ldots, s_6$. For distance d, we choose the Euclidean distance up to the standardization of the statistics components. Then, we obtain six posterior samples $\Theta_{\epsilon,I}^{(j)}$, $j = 1, \ldots, 6$. These computations are repeated K times and each time a new observed data set $\{(\theta^{(k)}, \mathcal{D}^{(k)}) : k = 1, \ldots, K\}$ is used. Three indices characterizing the posterior samples are assessed for each parameter:

- Posterior interval (PI) coverage : mean percentage of 95%-PIs which include the true value;
- PI length: mean length of the marginal 95%-PIs;
- Mean Square Error (MSE): MSE calculated with the marginal posterior median.

Parameter μ

Statistics	PI coverage (%)	PI length	Mean Square Error (*100)
s_1 (minimal sufficiency)	95.2 (93.9,96.5)	1.17 (1.13,1.21)	11.62 (10.07,13.18)
s_2 (nonlinearity)	95.5 (94.2,96.8)	1.18 (1.14,1.22)	11.81 (10.21,13.41)
s_3 (redundancy)	95.8 (94.6,97.0)	1.16 (1.12,1.20)	11.46 (9.93,12.99)
s_4 (dimension increase)	96.9 (95.8,98.0)	1.22 (1.18,1.26)	11.44 (9.95,12.92)
s_5 (un-informativeness)	96.8 (95.7,97.9)	1.28 (1.24,1.31)	11.49 (9.98,12.99)
s_6 (all handicaps)	96.9 (95.8,98.0)	1.32 (1.28,1.36)	11.60 (10.06,13.14)

Parameter σ

Statistics	PI coverage (%)	PI length	Mean Square Error (*100)
s_1 (minimal sufficiency)	95.7 (94.4,97.0)	0.85 (0.83,0.88)	6.21 (5.47,6.95)
s_2 (nonlinearity)	95.1 (93.8,96.4)	0.88 (0.86,0.90)	6.39 (5.64,7.15)
s_3 (redundancy)	96.3 (95.1,97.5)	0.92 (0.90,0.95)	6.34 (5.56,7.11)
s_4 (dimension increase)	98.6 (97.9,99.3)	1.12 (1.09,1.15)	6.70 (5.90,7.51)
s_5 (un-informativeness)	96.5 (95.4,97.6)	1.00 (0.98,1.02)	6.40 (5.64,7.15)
s_6 (all handicaps)	98.3 (97.5,99.1)	1.24 (1.23,1.26)	8.51 (7.64,9.37)

Table 1: Characterization of the posterior distributions of parameters μ and σ obtained with the six statistics functions $s_1, \ldots, s_6$ and application of **A2** algorithm on the toy example with $I = 100,000$ simulations, $\tau = 0.001$ the rate of accepted simulations and $K = 10^3$ repetitions carried out. Between brackets:95%-confidence intervals of the mean values [± standard error].

Results: We set $I = 100,000$ simulations, $\tau = 0.001$ the rate of accepted simulations and $K = 10^3$ repetitions carried out (see Table 1).

Concerning the PI coverage, we expect that 95% of the 95%-PIs include the true value of the parameter. For the parameter μ, it is around 95% for statistics s_1, s_2 and s_3 (resp. 95.2%, 95.5% and 95.8%) and their 95% confidence intervals (CI) contain the value 95% (for exemple, CI=[93.9%;96.5%] for s_1). The PI coverage for statistics s_4, s_5 and s_6 are significantly higher than 95% (the lower bound of their CIs are higher than 95%). For σ, only two statistics, s_1 and s_2, are around 95%. The PI coverages for statistics $s_3, \ldots, s_6$ are higher than 95%, in the same way than for μ.

The marginal 95%-PIs length for μ is significantly lower in average for statistics $s_1, \ldots, s_3$ (lengths around 1.17) than for statistics s_5 and s_6 (lengths resp. 1.28 and 1.32): for example CI(s_6)=[1.28;1.36] whereas CI(s_1)=[1.13;1.21]. The statistic s_4 is between those two groups, its CI intersects with both groups. One notice the same kind of results for σ: the lower group s_1 and s_2 is significantly different from the group with higher PI length $s_3, \ldots, s_6$. The length of 95% interval for σ is lower than for μ: for example, resp. 0.85 and 1.17 for statistic s_1. In summary, the marginal 95%-posterior intervals for μ and σ obtained with $s_3, \ldots, s_6$ tend to be larger than those obtained with s_1 and tend to overreach the 95% confidence level, contrary to those obtained with s_1. In constrast, results obtained with statistic s_2 (nonlinearity) are very close to those obtained with s_1. In this example,

the nonlinearity in the statistics does not affect inference on parameters.

Comparing the CIs, the MSE for μ is no significantly different between the six statistics $s_1, \ldots, s_6$. It is minimum for s_4 (MSE=11.44) and maximum for s_2 (MSE=11.81). Concerning the parameter σ, the statistic s_6 is significantly higher than the others (the lower bound for s_6 is 7.64 whereas the maximum of the other upper bounds is 7.51 for s_4).

Finally, the summary statistics that included all handicaps achieved a bad performance with MSE 1.4 times larger for σ and wide 95% confidence interval (length=1.24, PI coverage=98.3%). These results show that analyzing only the mean square error (MSE) to decide if a set of statistics give a good inference on parameters may be misleading, for parameter μ in this case: comparing MSE between statistics s_1 and s_6 for μ tends to choose s_6, but difference between both MSE is not significative and PI length is significantly larger for s_6 than s_1. The MSE criterion has to be completed with another criterion, like the length of the posterior interval, to analyze the quality of inference of a set of statistics.

Obviously, these results depend on the strength of the nonlinearity, the redundancy, the increase in the dimension and the un-informativeness. They also depend on I and τ: larger I and smaller τ tend to attenuate the differences between the minimal sufficient statistics and the other statistics. If increasing the number I of simulations is not possible, then a work concerning the statistics can be a fruitfull alternative. The approach proposed below is in this vein.

3 A new ABC procedure conditional on parameter point estimates

3.1 Method

Here, we propose to transform the statistics into point estimates of the parameters. These estimates are obtained by introducing and fitting a second and empirical model where the response variables are the parameters and the explanatory variables are the statistics:

$$\theta = l(S) + \xi,$$

where l is in a set of regression functions $\mathbb{L}$ which has to be specified and ξ is an error term. In a first step, based on a first set of simulated data, the empirical model is fitted. Then, for any statistics S we are able to estimate θ. The estimates $\hat{\theta}$ then replace the statistics S in the classical ABC procedure.

The aim of the empirical model is not to be informative, only to get point estimates that we could use as new statistics. The function $\hat{l} \in \mathbb{L}$ is chosen to minimize a discrepancy criterion between the simulated parameters θ and the transformed statistics $l(S)$ (see Section 3.4 for the choice of the discrepancy criterion and the set of functions $\mathbb{L}$).

The empirical model can be calibrated with the following algorithm:

A0. Carry out the next three steps

- For j in $\{1, \ldots, J\}$, independently generate θ_j from π and $\mathcal{D}_j$ from $\mathcal{M}_{\theta_j}$;
- Compute the statistics $S_j = s(\mathcal{D}_j)$, $j = 1, \ldots, J$;
- Estimate the function $\hat{l} \in \mathbb{L}$ which minimizes the discrepancy between the simulated parameters θ_j and the transformed statistics $l(S_j)$, $j = 1, \ldots, J$, $l \in \mathbb{L}$.

3.2 Incorporating the parameter point estimates in ABC–rejection

Incorporating the proposed transformation into algorithm **A2** provides the following algorithm:

A2*. 1. Estimate empirical model $\hat{l}$ with algorithm **A0**

2. Carry out the next three steps, independently for i in $\{1, \ldots, I\}$:
   - Generate θ_i from π and simulate $\mathcal{D}_i$ from $\mathcal{M}_{\theta_i}$;
   - Compute the statistics $S_i = s(\mathcal{D}_i)$, then compute the point estimates $\hat{\theta}_i = \hat{l}(S_i)$ of θ_i;
   - Accept θ_i if $d_\Theta(\hat{\theta}_i, \hat{\theta}) \leq \epsilon_\Theta$, where d_Θ is a distance over Θ and ϵ_Θ is a tolerance threshold for the distance between the point estimate $\hat{\theta} = \hat{l}(S)$ and the point estimates of the simulated parameters.

The set of accepted parameters, say $\tilde{\Theta}_{\epsilon,I} = \{\theta_i : d_\Theta(\hat{\theta}_i, \hat{\theta}) \leq \epsilon_\Theta, i = 1, \ldots, I\}$, forms a sample from the posterior distribution

$$p_{\epsilon_\Theta}(\theta \mid \hat{\theta}) = \frac{\left(\int_{B(\hat{\theta}, \epsilon_\Theta)} h(z \mid \theta) dz\right) \pi(\theta)}{\int_\Theta \left(\int_{B(\hat{\theta}, \epsilon_\Theta)} h(z \mid \alpha) dz\right) \pi(\alpha) d\alpha},$$

where $h(\hat{\theta} \mid \theta)$ is the conditional probability distribution function of $\hat{\theta}$ given θ.

3.3 Incorporating the parameter point estimates in ABC–Beaumont

Classical ABC–Beaumont can be described as follows:

A3. 1. Carry out the next three steps, independently for i in $\{1, \ldots, I\}$,

- Generate θ_i from π and simulate $\mathcal{D}_i$ from $\mathcal{M}_{\theta_i}$.
- Compute the statistics $S_i = s(\mathcal{D}_i)$, then compute the distance $d(S_i, S)$.
- Weight parameter θ_i according to the value of the distance using a kernel function $K_\delta(d(S_i, S))$, where $K_\delta(.)$ is any kernel function with bandwidth δ.

2. Adjust a weighted local linear regression $\hat{r}$ of θ_i on S_i and estimate the adjusted values θ_i^* of θ_i as: $\theta_i^* = \hat{r}(S) + \hat{\eta}_i$, where $\hat{r}(S)$ is the point estimate on observed statistic S and $\hat{\eta}_i = \theta_i - \hat{r}(S_i)$ the residuals of the regression.
3. Approximate the posterior density $\hat{p}$ of θ^* on θ using a kernel density estimation applied to the weighted sample:

$$\hat{p}(\theta \mid S) = \frac{\sum_i K_\Delta(d_\Theta(\theta_i^*, \theta)) K_\delta(d(S_i, S))}{\sum_i K_\delta(d(S_i, S))}$$

where $K_\Delta(.)$ is any kernel function with bandwidth Δ.

Asymptotic properties of this estimator are given in Blum (2010) and Hyndman et al. (1996).

Incorporating the proposed transformation into algorithm **A3** provides the following algorithm:

A3*. 1. Estimate empirical model $\hat{l}$ with algorithm **A0**
2. Carry out the next three steps, independently for i in $\{1, \ldots, I\}$,
   - Generate θ_i from π and simulate $\mathcal{D}_i$ from $\mathcal{M}_{\theta_i}$.
   - Compute the statistics $S_i = s(\mathcal{D}_i)$, then compute the point estimates $\hat{\theta}_i = \hat{l}(S_i)$ of θ_i and the distance $d_\Theta(\hat{\theta}_i, \hat{\theta})$.
   - Weight parameter θ_i according to the value of the distance using a kernel function $K_\delta(d_\Theta(\hat{\theta}_i, \hat{\theta}))$, where $K_\delta(.)$ is any kernel function with bandwidth δ.

3. Adjust a weighted local linear regression $\hat{r}$ of θ_i on $\hat{\theta}_i$ and estimate the adjusted values θ_i^* of θ_i as: $\theta_i^* = \hat{r}(\hat{\theta}) + \hat{\eta}_i$, where $\hat{r}(\hat{\theta})$ is the point estimate on observed and transformed statistic $\hat{\theta}$ and $\hat{\eta}_i = \theta_i - \hat{r}(\hat{\theta}_i)$ the residuals of the regression.
3. Approximate the posterior density $\hat{p}$ of θ^* on θ using a kernel density estimation applied to the weighted sample:

$$\hat{p}(\theta \mid \hat{\theta}) = \frac{\sum_i K_\Delta(d_\Theta(\theta_i^*, \theta)) K_\delta(d_\Theta(\hat{\theta}_i, \hat{\theta}))}{\sum_i K_\delta(d_\Theta(\hat{\theta}_i, \hat{\theta}))}$$

where $K_\Delta(.)$ is any kernel function with bandwidth Δ.

One could think that the kernel $K_\Delta(.)$ should capture the information and that our correction should not improve the estimator but the estimates $\hat{\theta}$ used as new statistics could be more correlated with θ than the raw statistics S. Consequently, the result of the local linear regression of θ_i on $\hat{\theta}_i$ could be more accurate than the one on S_i.

3.4 Choice of the set of regression functions $\mathbb{L}$ and the discrepancy criterion

The aim of the empirical model is not to be explicative of the mechanisms that provided the observations but to get point estimates of parameters.

Class of functions $\mathbb{L}$: we choose the set of *projection pursuit regressions*, or *PPR* (Friedman and Stuetzle, 1981; Hall, 1989). It presents some advantages compared with methods previously used (Joyce and Marjoram, 2008; Wegmann et al., 2009; Blum and François, 2010; Fearnhead and Prangle, 2010): (i) PPR allows to fit a different model for each component of parameter θ, (ii) a PPR model is a linear combination of smoothed linear or nonlinear functions of explanatory variables that naturally considers interactions of predictors, (iii) PPR does not require specification of a metric in the predictor space, (iv) PPR can estimate a much larger class of models than a simple linear regression.
Each component of the parameter θ is explained separately using all summary statistics as explanatory variables. The set of PPR functions is defined by Friedman and Stuetzle (1981) as:

$$l(S) = \sum_{m=1}^{M} f_{\alpha_m}(\alpha_m.S)$$

where α_m is a vector of coefficient, $\alpha_m.S$ the inner product, f_{α_m} a univariate function and M the number of functions of explanatory variables to be used.

Considering a parameter to explain $\theta^{[k]} \in \mathbb{R}$ and the explanatory variables $S \in \mathbb{R}^p$, the model is fitted iteratively by the following algorithm:

1. Set $r_i^{[0]} = \theta_i^{[k]}$
2. For $j = 1, \ldots$ maximize

$$R_{[j]} = 1 - \frac{\sum_{i=1}^{I} \left(r_i^{[j-1]} - f_{[j]}^{[k]}(\alpha_{[j]}^{[k]}.S_i) \right)^2}{\sum_{i=1}^{I} \left(r_i^{[j-1]} \right)^2}$$

   fitting the coefficients $\alpha_{[j]}^{[k]} \in \mathbb{R}^p$ and a univariate regression function $f_{[j]}^{[k]}$.
3. Define $r_i^{[j]} = r_i^{[j-1]} - f_{[j]}^{[k]}(\alpha_{[j]}^{[k]}.S_i)$ and repeat step 2 until $R_{[j]}$ is small.

Depending on the software, a threshold for a small $R_{[j]}$ can be defined to stop the algorithm. We choose to use the *ppr* function of the software *R* with default options, only precising the *nterms* option to the total number of summary statistics used to fit the model.

Discrepancy criterion: we choose the least squares as discrepancy criterion between the simulated parameters $\theta^{[k]}$ and the transformed statistics $l^{[k]}(S)$:

$$\hat{l}^{[k]} = \underset{l^{[k]} \in \mathbb{L}}{\arg\min} \sum_{i=1}^{I} \left(\theta_i^{[k]} - l^{[k]}(S_i) \right)^2$$

which has to be minimized in $\mathbb{L}$ a set of regression functions.

4 Application to a toy example: parameters of a Gaussian distribution

We study the performance of our estimator on the same gaussian example as presented in Section 2, using the same model and the same statistics $s_j, j = 1, \ldots, 6$. We apply the algorithms A2 (Rejection), A2* (PPR + Rejection), A3 (Beaumont) and A3* (PPR + Beaumont).

Parameter μ

	Stat.	A2	A2*	A3	A3*
Bias (*100)	s_1	0.23 (-1.88,2.35)	0.05 (-2.07,2.17)	0.30 (-1.82,2.42)	0.23 (-1.90,2.35)
	s_6	0.35 (-1.77,2.46)	0.11 (-2.01,2.23)	0.33 (-1.79,2.46)	0.07 (-2.06,2.21)
MSE (*100)	s_1	11.62 (10.07,13.18)	11.67 (10.11,13.23)	11.69 (10.11,13.28)	11.76 (10.18,13.33)
	s_6	11.60 (10.06,13.14)	11.67 (10.13,13.21)	11.71 (10.13,13.29)	11.78 (10.22,13.33)
PI cov. (%)	s_1	95.2 (93.9,96.5)	95.4 (94.1,96.7)	94.5 (93.1,95.9)	94.5 (93.1,95.9)
	s_6	96.9 (95.8,98.0)	95.1 (93.8,96.4)	96.3 (95.1,97.5)	94.6 (93.2,96.0)
PI length	s_1	1.17 (1.13,1.21)	1.17 (1.13,1.21)	1.17 (1.13,1.21)	1.16 (1.12,1.20)
	s_6	1.32 (1.28,1.36)	1.17 (1.13,1.21)	1.27 (1.23,1.31)	1.17 (1.13,1.21)

Parameter σ

	Stat.	A2	A2*	A3	A3*
Bias (*100)	s_1	-3.07 (-4.60, -1.53)	-3.07 (-4.62,-1.53)	-3.12 (-4.66,-1.59)	-3.08 (-4.62,-1.53)
	s_6	-12.44 (-14.08,-10.81)	-3.08 (-4.62,-1.53)	-10.57 (-12.19,-8.96)	-3.09 (-4.63,-1.54)
MSE (*100)	s_1	6.21 (5.47,6.95)	6.27 (5.52,7.03)	6.22 (5.47,6.97)	6.28 (5.52,7.05)
	s_6	8.51 (7.64,9.37)	6.27 (5.52,7.03)	7.88 (7.04,8.72)	6.28 (5.53,7.04)
PI cov. (%)	s_1	95.7 (94.4,97.0)	95.5 (94.2,96.8)	95.2 (93.9,96.5)	95.4 (94.1,96.7)
	s_6	98.3 (97.5,99.1)	95.1 (93.8,96.4)	97.9 (97.0,98.8)	94.6 (93.2,96.0)
PI length	s_1	0.85 (0.83,0.88)	0.85 (0.82,0.87)	0.84 (0.82,0.87)	0.84 (0.82,0.87)
	s_6	1.24 (1.23,1.26)	0.84 (0.82,0.87)	1.18 (1.16,1.20)	0.84 (0.81,0.86)

Table 2: Characterization of the posterior distributions of parameters μ and σ obtained with statistics functions s_1 and s_6 and application of **A2** (ABC–Rejection), **A2*** (PPR + ABC–Rejection), **A3** (ABC–Beaumont) and **A3*** (PPR + ABC–Beaumont) algorithms on the toy example with $I = 100,000$ simulations, $\tau = 0.001$ the rate of accepted simulations and $K = 10^3$ repetitions carried out. Between brackets:95%-confidence intervals of the mean values.

Bias, Mean Square Error, PI coverage and PI length: We expect that inferences made with statistics corrected by a PPR regression model will provide better results thant inferences made by uncorrected statistics in the case of handicaped summary statistics (cases $s_2, \ldots, s_6$). We also investigate wether statistics corrected by PPR regression model succeed in reaching similar accuracy than the non-corrected sufficient statistics.

We only present minimal sufficient statistics s_1 in comparison to statistic gathering all handicaps s_6 (for exhaustive results on statistics $s_j, j = 1, \ldots, 6$ see Appendix A, Tables 3

and 4). Concerning the parameter μ, the bias and the MSE are not significantly different between all algorithms and statistics: for example, with alogorith A3, MSE for s_1 is 11.69 (CI=[10.11,13.28]) and is 11.71 for s_6 (CI=[10.13,13.29]) (see Table 2). On the other hand, the PI coverage and the PI length, for each pair of algorithms (A2 and A2* or A3 and A3*), are significantly different between sufficient statistics and uncorrected handicap statistics when it is not between sufficient statistics and handicap statistics corrected by PPR model regression. The PI coverage is higher than 95% for uncorrected s_6 with algorithm A2 (CI=[95.8,98.0]) when it contains the 95% value for s_1 (CI=[93.9,96.5]) and for corrected s_6 with A2* algorithm (CI=[93.8,96.4]). In the same way, the PI length for uncorrected s_6 is significantly larger than those for s_1 or corrected s_6 in ABC–Rejection or in ABC–Beaumont.

Concerning the parameter σ, all results have the same profile: corrected handicap statistics succeed to reach same accuracy than non-corrected sufficient statistics when the non-corrected handicap statistics does not. For example in ABC–Beaumont (algorithms A3 and A3*), the bias for non-corrected s_6 (bias=-10.57) is larger in absolute value than the one for non-corrected s_1 (bias=-3.12) or corrected s_6 (bias=-3.09); also, the MSE, the PI coverage and the PI length are larger for non-corrected s_6 than for non-corrected s_1 and corrected s_6. The results for ABC–Rejection lead to the same conclusions.

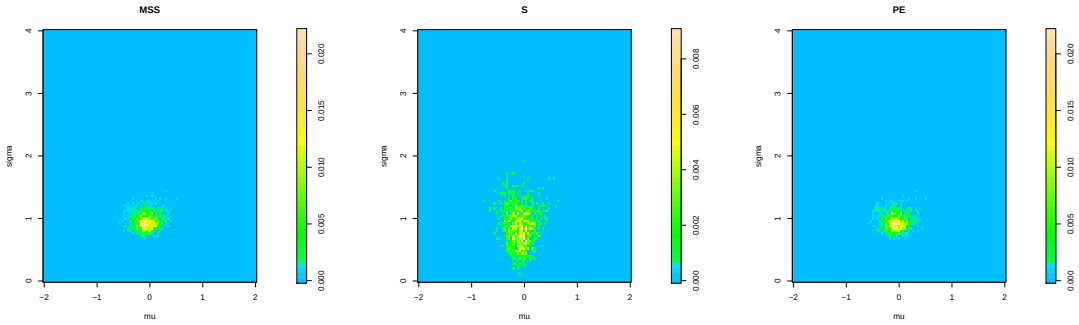
On this example, the correction allows to reach the expected results, except for the bias and MSE on the parameter μ. Also, as already shown in Section 2, the MSE criterion has to be completed with the length of the posterior interval to analyze the quality of inference of the sets of statistics.

Estimation of the posterior density: Figure 1 shows the effect of the PPR correction on the estimation of the joint posterior density of parameters (μ, σ) on a particular example. We set $I = 1,000,000$ simulations and $\tau = 0.001$ of accepted parameters and used algorithms A2 on s_1 and s_6 and A2* on s_6.

The area of the posterior density estimated with point estimates s_6^* is as reduced as the one obtained for minimal sufficient statistics s_1. Contrastedly, the one estimated with statistic s_6 is more extended.

5 Discussion

Approximate Bayesian computation procedures have been developped to allow estimation of parameters when the likelihood of the data is not tractable, such as the case of implicit statistical models. They are based on summary statistics and a tolerance threshold and face the same difficulty: the choice of the summary statistics. This choice can be more or less difficult according to the field of analyze and the knowledge of the relation between available summary statistics and parameters to estimate. When the relation is not well known, the analyst has to be helped to understand the role of summary statistics, to choose



Minimal Sufficient Statistics | Handicaped Statistics | Point Estimates (by PPR)

Figure 1: Estimation of the joint density by ABC-rejection made on 10^6 simulations with 10^{-3} of accepted parameters for $(\mu, \sigma) = (0.05, 1.15)$

them or decide to let them all, and to correct them before running an ABC procedure if it is needed.

The toy example shows that the dimension of the statistics and their handicaps (non-linearity, redundancy, un-informativeness, high dimensionality) are problems that are not managed correctly by classical ABC procedures. Each handicap does not play the same role: the nonlinearity represented by statistic s_2 lead to results close to those of sufficient statistic s_1. The handicaps do not cumulate together: results for the statistic s_6 gathering all handicaps is not the addition of the results for the statistics $s_2, \ldots, s_5$ representing each of the handicaps. The choice of a set of summary statistics can be based on the MSE criterion (Beaumont et al., 2002; Joyce and Marjoram, 2008). The calculation of the MSE for the parameter μ on the toy example makes appear that it can not be taken as only criterion to decide of the quality of inference of a set of summary statistics. The MSE completed with another criterion like the length of posterior interval can be used as a tool of validation of the proposed transformation of the statistics. The toy example also allows to compare the ABC–Rejection and ABC–Beaumont procedures and shows that, besides some few cases, the ABC–Beaumont procedure leads to results closer to those of sufficient statistics than the ABC–Rejection procedure.

Very recently, Fearnhead and Prangle (2010) propose to construct summary statistics choosing posterior means of the parameters. The point estimates we propose are in the same vein, using a PPR regression model for the transformation of data. In the toy example, these point estimates are as efficient as sufficient statistics, even for summary statistics gathering all handicaps.

The proposed method presents two advantages. Concerning the computation time, the model is estimated in a first step, independent of the ABC procedure which comes in a second step. The calibration of the model does not require a large set of data: for

example, we used 10^4 simulations to calibrate the 12 models of the toy example (6 sets of summary statistics x 2 parameters). The independance between the calibration and the ABC procedure makes that this method of transformation can be integrated and used in any ABC procedure (ABC–Beaumont, ABC–Blum, ABC–MCMC, ABC–PMC, ABC–SMC,...).

It could be interesting to replace the PPR regression model by another set of regression models, like linear models for example, and see result on the accuracy of the transformation. In the same way, the example we used is a very simple one in which the sufficient statistics are known. An application on a more difficult example, like a coalescent model, would be a good future improvement.

A Appendix

Bias for μ (*100)				
Statistics	Rejection (A2)	PPR + Rejection (A2*)	Beaumont (A3)	PPR + Beaumont (A3*)
s_1 (minimal sufficiency)	0.23 (-1.88,2.35)	0.05 (-2.07,2.17)	0.30 (-1.82,2.42)	0.23 (-1.90,2.35)
s_2 (nonlinearity)	-0.07 (-2.21,2.06)	0.11 (-2.01,2.24)	0.08 (-2.05,2.21)	0.25 (-1.88,2.38)
s_3 (redundancy)	0.09 (-2.01,2.19)	0.13 (-2.00,2.26)	0.20 (-1.91,2.32)	0.21 (-1.92,2.33)
s_4 (dimension increase)	0.18 (-1.92,2.28)	0.24 (-1.87,2.35)	0.12 (-2.00,2.24)	0.22 (-1.91,2.34)
s_5 (un-informativeness)	-0.03 (-2.14,2.07)	0.11 (-2.01,2.22)	-0.11 (-2.22,2.00)	0.17 (-1.96,2.30)
s_6 (all handicaps)	0.35 (-1.77,2.46)	0.11 (-2.01,2.23)	0.33 (-1.79,2.46)	0.07 (-2.06,2.21)

Bias for σ (*100)				
Statistics	Rejection (A2)	PPR + Rejection (A2*)	Beaumont (A3)	PPR + Beaumont (A3*)
s_1 (minimal sufficiency)	-3.07 (-4.60, -1.53)	-3.07 (-4.62,-1.53)	-3.12 (-4.66,-1.59)	-3.08 (-4.62,-1.53)
s_2 (nonlinearity)	-3.23 (-4.79, -1.67)	-3.09 (-4.63,-1.55)	-3.20 (-4.76,-1.65)	-3.09 (-4.63,-1.55)
s_3 (redundancy)	-3.25 (-4.80, -1.70)	-3.20 (-4.75,-1.65)	-3.21 (-4.76,-1.66)	-3.18 (-4.73,-1.63)
s_4 (dimension increase)	-5.08 (-6.66, -3.51)	-3.05 (-4.59,-1.51)	-4.52 (-6.10,-2.95)	-2.99 (-4.53,-1.44)
s_5 (un-informativeness)	-3.38 (-4.94, -1.83)	-3.13 (-4.68,-1.58)	-3.32 (-4.88,-1.76)	-3.13 (-4.67,-1.59)
s_6 (all handicaps)	-12.44 (-14.08,-10.81)	-3.08 (-4.62,-1.53)	-10.57 (-12.19,-8.96)	-3.09 (-4.63,-1.54)

Mean Square Error for μ (*100)				
Statistics	Rejection (A2)	PPR + Rejection (A2*)	Beaumont (A3)	PPR + Beaumont (A3*)
s_1 (minimal sufficiency)	11.62 (10.07,13.18)	11.67 (10.11,13.23)	11.69 (10.11,13.28)	11.76 (10.18,13.33)
s_2 (nonlinearity)	11.81 (10.21,13.41)	11.71 (10.14,13.28)	11.76 (10.15,13.37)	11.78 (10.20,13.36)
s_3 (redundancy)	11.46 (9.93,12.99)	11.76 (10.20,13.33)	11.60 (10.05,13.16)	11.73 (10.18,13.29)
s_4 (dimension increase)	11.44 (9.95,12.92)	11.57 (10.02,13.13)	11.68 (10.13,13.22)	11.71 (10.10,13.33)
s_5 (un-informativeness)	11.49 (9.98,12.99)	11.62 (10.05,13.18)	11.53 (9.99,13.06)	11.77 (10.17,13.36)
s_6 (all handicaps)	11.60 (10.06,13.14)	11.67 (10.13,13.21)	11.71 (10.13,13.29)	11.78 (10.22,13.33)

Mean Square Error for σ (*100)				
Statistics	Rejection (A2)	PPR + Rejection (A2*)	Beaumont (A3)	PPR + Beaumont (A3*)
s_1 (minimal sufficiency)	6.21 (5.47,6.95)	6.27 (5.52,7.03)	6.22 (5.47,6.97)	6.28 (5.52,7.05)
s_2 (nonlinearity)	6.39 (5.64,7.15)	6.28 (5.53,7.03)	6.36 (5.61,7.11)	6.26 (5.50,7.02)
s_3 (redundancy)	6.34 (5.56,7.11)	6.32 (5.56,7.08)	6.34 (5.57,7.12)	6.34 (5.57,7.11)
s_4 (dimension increase)	6.70 (5.90,7.51)	6.26 (5.51,7.02)	6.64 (5.84,7.44)	6.29 (5.54,7.05)
s_5 (un-informativeness)	6.40 (5.64,7.15)	6.32 (5.55,7.09)	6.42 (5.65,7.19)	6.25 (5.49,7.01)
s_6 (all handicaps)	8.51 (7.64,9.37)	6.27 (5.52,7.03)	7.88 (7.04,8.72)	6.28 (5.53,7.04)

Table 3: Characterization of the mean bias and mean square error for μ and σ obtained with six statistics functions $s_1, \ldots, s_6$ and application of **A2, A2*, A3 and A3*** algorithms for a normal sample from $N(\mu, \sigma)$, $\mu \in [-2; 2]$ and $\sigma \in (0; 4]$. 10^3 repetitions where carried out. The Square Error for μ and σ is calculated with the posterior median. Between brackets:95%-confidence intervals of the mean values.

PI coverage for μ (%)				
Statistics	Rejection (A2)	PPR + Rejection (A2*)	Beaumont (A3)	PPR + Beaumont (A3*)
s_1 (minimal sufficiency)	95.2 (93.9,96.5)	95.4 (94.1,96.7)	94.5 (93.1,95.9)	94.5 (93.1,95.9)
s_2 (nonlinearity)	95.5 (94.2,96.8)	95.3 (94.0,96.6)	95.0 (93.6,96.4)	95.0 (93.6,96.4)
s_3 (redundancy)	95.8 (94.6,97.0)	95.6 (94.3,96.9)	95.2 (93.9,96.5)	95.1 (93.8,96.4)
s_4 (dimension increase)	96.9 (95.8,98.0)	94.8 (93.4,96.2)	95.5 (94.2,96.8)	94.7 (93.3,96.1)
s_5 (un-informativeness)	96.8 (95.7,97.9)	94.9 (93.5,96.3)	96.4 (95.2,97.6)	94.4 (93.0,95.8)
s_6 (all handicaps)	96.9 (95.8,98.0)	95.1 (93.8,96.4)	96.3 (95.1,97.5)	94.6 (93.2,96.0)

PI coverage for σ (%)				
Statistics	Rejection (A2)	PPR + Rejection (A2*)	Beaumont (A3)	PPR + Beaumont (A3*)
s_1 (minimal sufficiency)	95.7 (94.4,97.0)	95.5 (94.2,96.8)	95.2 (93.9,96.5)	95.4 (94.1,96.7)
s_2 (nonlinearity)	95.1 (93.8,96.4)	95.7 (94.4,97.0)	94.8 (93.4,96.2)	95.5 (94.2,96.8)
s_3 (redundancy)	96.3 (95.1,97.5)	95.3 (94.0,96.6)	95.6 (94.3,96.9)	95.6 (94.3,96.9)
s_4 (dimension increase)	98.6 (97.9,99.3)	95.1 (93.8,96.4)	98.2 (97.4,99.0)	94.8 (93.4,96.2)
s_5 (un-informativeness)	96.5 (95.4,97.6)	95.0 (93.6,96.4)	96.7 (95.6,97.8)	94.3 (92.9,95.7)
s_6 (all handicaps)	98.3 (97.5,99.1)	95.1 (93.8,96.4)	97.9 (97.0,98.8)	94.6 (93.2,96.0)

PI length for μ				
Statistics	Rejection (A2)	PPR + Rejection (A2*)	Beaumont (A3)	PPR + Beaumont (A3*)
s_1 (minimal sufficiency)	1.17 (1.13,1.21)	1.17 (1.13,1.21)	1.17 (1.13,1.21)	1.16 (1.12,1.20)
s_2 (nonlinearity)	1.18 (1.14,1.22)	1.16 (1.12,1.20)	1.17 (1.13,1.21)	1.16 (1.12,1.20)
s_3 (redundancy)	1.16 (1.12,1.20)	1.16 (1.13,1.20)	1.15 (1.11,1.19)	1.16 (1.12,1.20)
s_4 (dimension increase)	1.22 (1.18,1.26)	1.16 (1.12,1.20)	1.18 (1.14,1.22)	1.16 (1.12,1.20)
s_5 (un-informativeness)	1.28 (1.24,1.31)	1.17 (1.13,1.21)	1.24 (1.20,1.28)	1.16 (1.13,1.20)
s_6 (all handicaps)	1.32 (1.28,1.36)	1.17 (1.13,1.21)	1.27 (1.23,1.31)	1.17 (1.13,1.21)

PI length for σ				
Statistics	Rejection (A2)	PPR + Rejection (A2*)	Beaumont (A3)	PPR + Beaumont (A3*)
s_1 (minimal sufficiency)	0.85 (0.83,0.88)	0.85 (0.82,0.87)	0.84 (0.82,0.87)	0.84 (0.82,0.87)
s_2 (nonlinearity)	0.88 (0.86,0.90)	0.85 (0.83,0.88)	0.87 (0.84,0.89)	0.84 (0.82,0.87)
s_3 (redundancy)	0.92 (0.90,0.95)	0.85 (0.83,0.88)	0.90 (0.87,0.92)	0.85 (0.82,0.87)
s_4 (dimension increase)	1.12 (1.09,1.15)	0.85 (0.82,0.87)	1.05 (1.03,1.08)	0.84 (0.82,0.87)
s_5 (un-informativeness)	1.00 (0.98,1.02)	0.85 (0.83,0.88)	0.96 (0.94,0.98)	0.84 (0.82,0.87)
s_6 (all handicaps)	1.24 (1.23,1.26)	0.84 (0.82,0.87)	1.18 (1.16,1.20)	0.84 (0.81,0.86)

Table 4: Characterization of the means of PI length and PI coveragefor μ and σ obtained with six statistics functions $s_1, \ldots, s_6$ and application of **A2, A2*, A3 and A3*** algorithms for a normal sample from $N(\mu, \sigma)$, $\mu \in [-2; 2]$ and $\sigma \in (0; 4]$. 10^3 repetitions where carried out. The posterior interval (PI) coverage for μ (resp. σ) is the mean percentage of 95%-PIs which include the true value μ_0 (resp. σ_0). The PI lengths for μ and σ are the mean lengths of the marginal 95%-PIs. Between brackets:95%-confidence intervals of the percentages and mean values.

References

Beaumont, M., J. Marin, and J. Cornuet (2009). Adaptivity for abc algorithms: the abc-pmc. *Biometrika 96*(4), 983–990.

Beaumont, M., W. Zhang, and D. Balding (2002). Approximate bayesian computation in population genetics. *Genetics 162*(4), 2025.

Blum, M. and O. François (2010). Non-linear regression models for approximate bayesian computation. *Statistics and Computing 20*(1), 63–73.

Blum, M. G. B. (2010, SEP). Approximate bayesian computation: A nonparametric perspective. *Journal of the American statistical Association 105*(491), 1178–1187.

Diggle, P. and R. Gratton (1984). Monte carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological) 46*(2), 193–227.

Estoup, A., M. Beaumont, F. Sennedot, C. Moritz, and J. Cornuet (2004). Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, bufo marinus. *Evolution 58*(9), 2021–2036.

Fearnhead, P. and D. Prangle (2010). Semi-automatic approximate bayesian computation. *Arxiv preprint arXiv:1004.1112*.

Friedman, J. and W. Stuetzle (1981). Projection pursuit regression. *Journal of the American statistical Association 76*(376), 817–823.

Guillemaud, T., M. Beaumont, M. Ciosi, J. Cornuet, and A. Estoup (2009). Inferring introduction routes of invasive species using approximate bayesian computation on microsatellite data. *Heredity 104*(1), 88–99.

Hall, P. (1989). On projection pursuit regression. *The Annals of Statistics 17*(2), 573–588.

Hamilton, G., M. Currat, N. Ray, G. Heckel, M. Beaumont, and L. Excoffier (2005). Bayesian estimation of recent migration rates after a spatial expansion. *Genetics 170*(1), 409.

Hyndman, R. J., D. M. Bashtannyk, and G. K. Grunwald (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics 5*(4), pp. 315–336.

Joyce, P. and P. Marjoram (2008). Approximately sufficient statistics and bayesian computation. *Statistical Applications in Genetics and Molecular Biology 7*(1), 26.

Marin, J., P. Pudlo, C. Robert, and R. Ryder (2011). Approximate bayesian computational methods. *Arxiv preprint arXiv:1101.0955*.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003). Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America 100*(26), 15324.

Nunes, M. A. and D. J. Balding (2010). On optimal selection of summary statistics for approximate bayesian computation. *Statistical Applications in Genetics and Molecular Biology 9*(1).

Ovaskainen, O. and I. Hanski (2004). Metapopulation dynamics in highly fragmented landscapes. In *Ecology, Genetics, and Evolution of Metapopulations*, Chapter 4, pp. 73–103. Elsevier.

Pritchard, J., M. Seielstad, A. Perez-Lezaun, and M. Feldman (1999). Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular Biology and Evolution 16*(12), 1791.

Real, L. and R. Biek (2007). Spatial dynamics and genetics of infectious diseases on heterogeneous landscapes. *Journal of the Royal Society Interface 4*(16), 935.

Rubin, D. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics 12*(4), 1151–1172.

Sisson, S. A., Y. Fan, and M. M. Tanaka (2007). Sequential monte carlo without likelihoods. *104*(6), 1760–1765.

Soubeyrand, S., A. Laine, I. Hanski, and A. Penttinen (2009). Spatiotemporal structure of host-pathogen interactions in a metapopulation. *The American Naturalist 174*(3), 308–320.

Wegmann, D., C. Leuenberger, and L. Excoffier (2009). Efficient approximate bayesian computation coupled with markov chain monte carlo without likelihood. *Genetics 182*(4), 1207.

Wilkinson, R. (2008). Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Arxiv preprint arXiv:0811.3355*.