MOLECULAR ECOLOGY
RESOURCES

Kernel-beta: A novel non-parametric method for model choice in approximate Bayesian computation

Journal:	*Molecular Ecology Resources*
Manuscript ID:	MER-14-0268
Manuscript Type:	Resource Article
Date Submitted by the Author:	03-Jul-2014
Complete List of Authors:	Sandoval-Castellanos, Edson; Swedish Museum of Natural History, Department of Bioinformatics and Genetics
Keywords:	Approximate Bayesian computation, Model choice, Bayesian estimation, Regression, Statistics, Hypothesis contrast

Kernel-beta: A novel non-parametric method for model choice in approximate Bayesian computation

Edson Sandoval-Castellanos[1]

[1] Department of Bioinformatics and Genetics, Swedish Museum of Natural History, Stockholm, 104 05, Sweden

Contact:

Edson Sandoval-Castellanos

Margaritas 60, San Pedro Martir II, Tlalpan

14640, Mexico City, Mexico

Tel. (+52)(1) 55 3877 1715

e-mail: EdsonSandovalC@outlook.com

Running title: Kernel-beta for Bayesian model choice

Subject area: Molecular and Statistical Advances

Abstract

The model choice is a critical exercise in science. It is performed increasingly by statistical methods that employ computer simulations due to their capacity to deal with complex problems. The model choice by approximate Bayesian computation (ABC) has many advantages due to the absence of likelihood estimation steps. However, the available methods for performing a model choice by ABC are few, and have known or potential limitations that can be exacerbated by high dimensionality. Here I present a novel non-parametric method that can circumvent other methods limitations by integrating an assumptions-free kernel regression with a Bayesian estimation of proportions in order to estimate model likelihoods and Bayes factors. This kernel-beta method altogether with a direct approach and a logistic regression, were comparatively tested with five simulated examples and one real dataset.

The new kernel-beta method consistently presented a lower error rate with only marginal loses of statistical power when compared to the direct approach, while the logistic regression presented disparate results that turned out to be caused by numerical issues. Conservative behavior could make the kernel-beta an optimal reference when different methods showed contrasting results or whit high dimensionality problems. In addition, the kernel-beta, the only method with a Bayesian fundament, can exploit the advantages of Bayesian inference and regression methods. Future developments should continue exploiting regression methods and look for fitting specific types of regression to different types of studies.

Introduction

The model choice, the adoption of one model as the best explanation of a natural phenomenon among a set of alternative models, is one of the most fundamental and critical exercises in science. It is performed predominantly by means of statistical methods (by hypothesis contrast) and, increasingly, by those that employ computer simulations.

In ecology and evolution, methods employing computer simulations are highly appreciated due to their capacity to deal with complex and multivariate problems (Hoban et al. 2012). Approximate Bayesian computation (ABC), also known as likelihood-free inference, is one of the most successful family of techniques due to its simplicity and flexibility, advantages that are granted by the absence of likelihood estimation steps (Beaumont 2010; Beaumont & Rannala 2004; Beaumont et al. 2002). ABC has been applied to the estimation of parameters as mutation, migration and selection rates (Pritchard et al. 1999; Hamilton et al. 2005; Jensen et al. 2008); ratios of population admixture (Sousa et al. 2009); times to population divergence (Lagerholm et al. 2014); and pathogens spread rates (Tanaka et al. 2006). ABC has also been applied to test alternative models of historical demography, migration, and population structure (Fagundes et al. 2007; Sjodin et al. 2012; Palkopoulou et al. 2013); as well as colonization (Ilves et al. 2010), species introgression (Roux et al. 2013), and domestication (Warmuth et al. 2012).

ABC consist in performing a large number of simulations that are accepted or rejected on the basis of their similarity with the observed data, as measured from a set of summary statistics (Beaumont et al. 2002). The unknown parameters that are required for the simulations are drawn from probability distributions which are considered Bayesian priors (Beaumont et al. 2002). The parameters values in the accepted simulations approach a sample of the Bayesian posterior of the parameters conditional to the observed data (Beaumont 2010). Model choice analyses, can be performed by means of Bayes factors that are obtained as the quotients between model likelihoods, which are the probabilities of observing the data

conditioned to specific models. They are also approached from the set of accepted simulations (Beaumont 2010).

ABC has received a number of improvements (see Beaumont 2010, for a review) targeting the improvement of accuracy and validation of results (Beaumont et al. 2002; Beaumont 2010; Bertorelle et al. 2010; Csillery et al. 2010); the problem of fully capturing the information contained in the data by means of the summary statistics –or Bayes sufficiency- (Beaumont 2010; Robert et al. 2011); and the computational efficiency (Beaumont et al. 2002; Marjoram et al. 2003; Sisson et al. 2007; Wegmann et al., 2009; Blum & Francois 2010; Leuenberger & Wegmann 2010). Model choice has also received some attention for improvement (Ratmann et al. 2009; Robert et al. 2011). However, there is still a sparse number of alternatives for performing it, namely: a direct approach, where acceptance proportions are directly operated as model likelihoods (Beaumont 2010); and the regression approaches of Leuenberger & Wegmann (2010) and Fagundes et al. (2007), where the model likelihoods are estimated from a relationship with the summary statistics that is assumed to be linear at a local scale, i.e. in a neighborhood of the observed values. Regression approaches have proved to increase the rate of right model choice (Leuenberger & Wegmann 2010; Fagundes et al. 2007), and are solid statistical procedures that can provide measures of error for the estimates of model likelihoods and Bayes factors. In addition, the information of distances between simulated and observed summary statistics is naturally incorporated into the estimation procedure. This feature is always desirable for ABC since the inference is not strictly conditioned to the observed data but to the accepted set. So, taking the information of the distances into account serves for weighting more the accepted simulations as their summary statistics get closer to the observed ones. The potential disadvantages of regression methods include the introduction of additional parameters and assumptions as those regarding the distribution of error and the linearity of the relationships among summary statistics, parameters and model likelihoods. Since regression is performed locally, it is considered that departures of linearity should be irrelevant. The problem comes when the

number of summary statistics and/or parameters is large, because the high dimensionality requires that the acceptance region be stretched in order to keep the computational effort inside the budget (Leuenberger & Wegmann 2010). This "curse of dimensionality", as it has been called, can make the acceptance region unexpectedly large and leading to relevant departures between the linear (flat) hyper-surface assumed by regression and the real hyper-surface of the parametric spaces of summary statistics and parameters, which can lead to important biases in the estimates obtained by regression.

Here I present a non-parametric Bayesian method to perform model choice analysis by means of approximate Bayesian computation. It could be robust to high dimensionality problems due to the lack of parameters and assumptions and incorporates the information of the distances among observed and simulated data. This method was also put on test by four simulated examples and one real dataset.

Materials and Methods

Fundament and rationale

The regression methods are powerful and mathematically solid. Since they can greatly improve the accuracy of the parameters estimation and model choice, it could be wise to keep some of their spirit while throwing off parameters and assumptions in order to dodge dimensionality problems. So, in first instance I pursuit a non-parametric and assumptions-free method with a regression-like fundament, as a complement, rather than a substitute to other regression methods. Secondly, recalling that regression methods (as implemented in ABC) are frequentist, it would be advantageous for a new approach to have a Bayesian fundament. A Bayesian fundament would provide a better error assessment, enable the possible use of prior information, and match the Bayesian nature of the ABC framework of which the model choice analysis is part.

So, the method described below is the result of integrating two concepts: a kernel regression, in which the Euclidean distances are used to obtain weighted acceptance ratios that approach the expectations of the model likelihoods without employing any parameters or assumptions; and a Bayesian estimation of proportions from a multinomial sample.

Model choice in approximate Bayesian computation

As a Bayesian method, ABC addresses a model choice problem by means of Bayes factors (BF). A BF is interpreted as the relative support that one model (M_i) gets respect another model (M_j) given the observed data, and is given by the quotient between the respective model likelihoods (Gelman et al., 2003). The magnitude of the BF is the criterion to pick one model over another one; for instance, a BF above 10 is considered to provide strong support to the model in the numerator. In the direct approach the model likelihoods are equivalent to the acceptance ratios (p_i, p_j):

$$BF(M_i; M_j) = \frac{f(x|M_i)}{f(x|M_j)} = \frac{p_i}{p_j} \approx \frac{P(\|s - s'\| < \delta_\varepsilon|M_i)}{P(\|s - s'\| < \delta_\varepsilon|M_j)}$$

where x is the observed data, s and s' are the simulated and observed summary statistics, $f(x|M_i)$ is the model likelihood (of the i-th model), and δ_ε represents the acceptance distance (Beaumont et al. 2010).

Kernel regression

The estimation of model likelihoods can be seen as an admixture problem (Fagundes et al., 2007), where the model likelihoods are the proportions that the different models have on the point that correspond to the values of the observed summary statistics. The estimation of those proportions is routinely obtained by interpolating the proportions in the entire acceptance region (direct approach), or from the predictive

function estimated by linear regression or logistic regression (regression approaches). In the kernel regression proposed here those proportions are obtained by weighting each simulation in the acceptance region by a kernel function evaluated in the distance between the summary statistics of the simulation and the observed ones. The normalized vector where each element is the sum of weights of each specific model, approaches the vector of expectations of the model likelihoods for a large enough sample. So, the model likelihood of the *i*-th model is estimated by: $f(x|M_i) = \frac{D_i}{D_1+D_2+\cdots+D_m}$; $D_i = \sum_{r=1}^{n} K_\delta(\|s_r - s'\|)\mathbb{I}_{M_i}$, where *n* is the overall number of accepted simulations, $\mathbb{I}_{M_i}$is the indicator function which value is 1 if the simulation under analysis corresponds to the *i*-th model, and 0 otherwise; and $K_\delta(\|s_r - s'\|)$ is the kernel with bandwidth δ evaluated in the distance between the summary statistics of the *r*-th simulation (s_r) and the observed summary statistics (s').

Three sensitive choices have to be done here: the kernel function, an appropriate bandwidth for the kernel function, and a useful distance. The kernel function should be one that maximizes the weight if the distance is zero and decreases continuously as the distance increases. The Epanechnikov kernel has already some predilection in ABC procedures but other kernels, as the Gaussian kernel, can be used (Beaumont 2002). The bandwidth determines the minimum distance that will be assigned a non zero weight. In other words the width of the meaningful neighborhood of the focal point (which is the point corresponding the observed summary statistics in our case). So it is reasonable to use a bandwidth coinciding with the distance for rejection. Finally, the Euclidean distance is already the standard choice for measuring distances in ABC so it could be safely adopted here.

Bayesian estimation

The kernel weighting incorporates the distances information into the estimation, but fails to provide of error measures. Here, that part will be fulfilled by considering the set of accepted simulations a

multinomial sample with the multiple states being the models the simulations were run with. So the problem is a Bayesian estimation of proportions given a multinomial sample, where the posterior probability density function (pdf) of the model likelihoods vector is the conjugate of a multinomial distribution, namely a Dirichlet distribution (Gelman et al. 2003). Its parameters are given by the model likelihoods estimations obtained by the kernel regression. In a regular multinomial sampling the Dirichlet distribution already scales the accuracy of the multivariate posterior by the sampling size as the parameters are the counts of the observations falling in the different categories, so the larger those numbers the lower the variance of their posteriors. Here, the sample size is righteously assigned by equaling the Dirichlet parameters to the unnormalized estimations of the model likelihoods obtained by the kernel regression procedure. In those estimations the overall sample size would be much smaller than the original number of accepted simulations because they have been weighted by the kernel function, which is correct since the weighting means a reduction in the contribution of the simulations according with their distances and consequently the sample size should be reduced in the same amount. Notice that in a regular, not weighted, multinomial sample the parameters of the Dirichlet density would be integers, but in our case the weighting procedure will make those parameters real numbers, which is not a problem since the parameters of a Dirichlet density can be real.

Now, since the BF are calculated from individual model likelihoods it could be useful to estimate the model likelihoods independently rather than jointly. In that case the marginal distribution of a Dirichlet, a beta distribution can be used (Gelman et al. 2003). Notice that this procedure yields a probability density function (pdf), a beta density, instead of a punctual value, so we have to obtain the statistical expectation to have an estimation of a Bayes factor. The expectation is given by integration of the model likelihoods weighted by their probabilities:

$$\widehat{BF}(M_i; M_j) = E\left[\frac{p_i}{p_j}\right] = \int_{p_j}\int_{p_i} \frac{p_i B(p_i|\beta_{i1},\beta_{i2})}{p_j B(p_j|\beta_{j1},\beta_{j2})}\delta p_2 \delta p_1 = \int_{p_j}\int_{p_i} \frac{p_i}{p_j}\delta B(p_i|\beta_{i1},\beta_{i2})\delta B(p_j|\beta_{j1},\beta_{j2})$$

, were $B(p_i|\beta_{i1},\beta_{i2})$ is the beta pdf obtained for the *i*-th model, which has parameters given by $\beta_{.1} = n^{-1}\sum_{k=1}^{n} K_\delta(\|s_r - s'\|)$; and $\beta_{i2} = 1 - \beta_{i1}$. In that expression, *n* is the overall number of accepted simulations, and $K_\delta(\|s_r - s'\|)$ is the Epanechnikov kernel evaluated in the Euclidean distance between the summary statistics of the *r*-th simulation (s_r) and the observed summary statistics (s'). The solution is $\frac{\beta_{i1}}{\beta_{j1}}$, which is also the maximum likelihood estimator of the BF. Unfortunately, credible intervals cannot be obtained analytically since the integral of a beta pdf is improper. Thus, they have to be obtained by Monte Carlo procedures (see below).

This approach endows the estimation with a measure of statistical error given by a proper sample size that is adjusted according to the information that is retained after the kernel weighting. It is also robust since it does not require additional parameters or assumptions.

Algorithm

The algorithm for performing a model choice analysis has two variants, depending on the way the simulations are run: independently for each model, or jointly by simulating all the models in the same run. The algorithm where the simulations of each model are run independently, consist in the next steps:

1. Run simulations in order to obtain a reference table. By using proper software, perform a simple rejection under an ABC framework without any regression adjustment.
2. Define an appropriate tolerance δ_ε for the Epanechnikov kernel. The acceptance distance employed for rejection could be a reasonable choice.
3. Calculate the distance between observed and simulated summary statistics ($\|s - s'\|$) and the weight by evaluating the distance in the Epenechnikov or other kernell: $K_\delta(\|s - s'\|)$. Perform this for every accepted simulation.

4. Estimate the beta pdf by defining its parameters: $\beta_1 = \sum_{r=1}^{n} K_\delta(\|s_r - s'\|)$; and $\beta_2 = n - \beta_1$, where n is the number of non-rejected simulations.
5. Repeat steps 1-4 for models M_1, M_2, M_3...(obtaining the beta distributions $B(p_1)$, $B(p_2)$, $B(p_3)$...). The rejection threshold and kernel tolerance should be the same for all models.
6. The punctual estimation of a model likelihood is $p = \frac{\beta_1}{\beta_1+\beta_2}$ while the credible interval can be obtained numerically by simulating random deviates from the respective beta pdf, $B(p_i)$ (simulation of random deviates from a beta pdf is trivial). Obtain punctual and interval estimations of each model likelihood.
7. The punctual estimations of the Bayes factors can be obtained as the quotient of their model likelihoods estimations. Credible intervals of the Bayes factors can be obtained numerically simulating pairs of beta deviates (from their respective beta pdf's) and getting their quotients. Obtain Bayes factors for every pair of models.
8. Perform a model choice by using the Bayes factors as criteria.

If the simulations of m alternative models are run jointly, then only a single reference table will be obtained, having an additional column for model indicator variables. In such case the sample would be multinomial and its conjugate posterior a Dirichlet pdf were the m parameters would be estimated as if each one was the parameter β_1 in the binomial version. The Bayes factors could be then obtained from the marginal densities of the models being compared. The algorithm is:

1. Run simulations for all the models jointly with probabilities according to prior odds (equal probabilities if there is no prior information regarding the models). The obtained reference table should contain a column to identify each simulation's model. By using proper software perform rejection under an ABC framework without any regression adjustment.

2. Define an appropriate tolerance δ_ε for the Epanechnikov kernel. The acceptance distance employed for the rejection could be a reasonable choice.
3. Calculate the distance between observed and simulated summary statistics ($\|s - s'\|$) and the weight by evaluating the distance in the Epenechnikov or other kernell: $K_\delta(\|s - s'\|)$. Perform this for every non-rejected simulation.
4. Estimate the parameter D of one model as the sum of all the weight values obtained for the accepted simulations corresponding each model: $D_i = \sum_{r=1}^{n} K_\delta(\|s_r - s'\|)\mathbb{I}_{M_i}$. Repeat this step for every model.
5. Normalize the D parameters by dividing them by their sum: $D_i^* = \frac{D_i}{D_1+D_2+\cdots+D_k}$, where k is the number of models. The joint distribution of the model likelihoods is a Dirichlet pdf with parameters being the (unormalized) D parameters of the models.
6. The punctual estimation of a model likelihood is D_i^* while joint or marginal credible intervals can be obtained numerically by simulating random deviates from the Dirichlet or the marginal beta pdf's respectively. The parameters of the marginal beta pdf's are $\beta_1 = D_i$, $\beta_2 = n - \beta_1$. Obtain credible intervals of each model likelihood.
7. The punctual estimations of the Bayes factors can be obtained as the quotient of their model likelihoods estimates. Credible intervals of the Bayes factors can be obtained numerically simulating pairs of beta deviates (from their respective beta pdf's) and getting their quotients. Obtain Bayes factors for every pair of models.
8. Perform a model choice by using the Bayes factors as criteria.

Testing and Validation

Three methods, the kernel-beta method described above, a direct approach, and a logistic regression, were applied to five simulated examples and one real dataset. Each simulated example consisted in a set of models that were tested, one at a time, by using a reference table with 1,000 pseudo-observed datasets (PODs) employed as the real data. Independently, an "instrumental" reference table was obtained, containing 1,000,000 simulations of each model. Reference tables contained the summary statistics calculated from DNA alignments obtained from coalescent simulations that were run under the different testing models. See Figure 1 for a schematic representation of the simulated examples (see Supplementary Information 1.1-1.5 for the input files of the five simulated examples, and Supplementary Tables 1-4 with the specifics of the priors employed). Simulations and rejection were run in the software BaySICS (Sandoval-Castellanos et al. 2014). A complete model choice analysis was applied to every one of the 1,000 PODs. The procedure was repeated for every model being the real one, so one reference table of PODs was created for every model. Each analysis consisted in an ABC procedure where the rejection algorithm selected the 5,000 closest simulations to the pseudo-observed data. Four different algorithms were applied to perform a model choice:

1. Direct approach. Model likelihoods were estimated as the proportions obtained in the accepted simulations, and the model with the highest likelihood was recorded as the chosen one.
2. Logistic regression without error. The accepted simulations were used for performing a logistic regression following Fagundes et al. (2007) to estimate the model likelihoods. The model with the highest likelihood was recorded as the chosen one.
3. Logistic regression with error. A Monte Carlo procedure was employed for obtaining confidence intervals of the model likelihoods. In it, the regression parameters were simulated from normal distributions with means and standard deviations matching the punctual and error estimates obtained by least squares in the logistic regression estimation. Then 10,000 random deviates of the parameters were used to get an equal number of model likelihoods, which were used to approach

their confidence intervals as well as the ones of the Bayes factors among models. A model was recorded as the chosen one if, for every comparison against other model, the entire confidence interval of the Bayes factor was larger than one.

4. Kernel-beta. The estimation of model likelihoods as well as their credible intervals was performed as described above. A model was recorded as the chosen one if, for every comparison against other model, the entire credible interval of the Bayes factor was larger than one.

The number of right and wrong model choices that every method scored were recorded for estimating the error rate and statistical power of each method. For this analyses a custom software was programmed in Fortran 95 language by using Microsoft Visual Studio and Intel Fortran Composer XE 2011. Regression procedures and random number generation were carried out by the specialized subroutines LAPACK and the module 'random' (A. Miller, available at: http://www.Mathtools.net).

The real dataset consisted in 741 bp mitochondrial DNA sequences from 59 woolly mammoths with ages ranging from 3.6 to 61.6 thousand years before present (kya), from Wrangell Island, Siberia (Nystrom et al. 2010). They were used to test four models of alternative demography at four periods of time (>12 kya, 9-12 kya, 6-9 kya, and 3-6 kya) that preceded their final extinction (Figure 1, H-K). Other settings included: generation time = 15 years; mutation rate = 24.7% change per million years; transition/transversion bias = 0.979 and gamma parameter = 0.05. An overall of 1,000,000 simulations were carried out for each model. Prior distributions are shown in Supplementary Table S5. Summary statistics were calculated for three statistical groups coinciding with the time periods defined in the models (except the third period, 9-12 kya that had no samples). Samples sizes were 18, 14 and 27 for the periods 3-6 kya, 6-9 kya and >12 kya, respectively. The employed summary statistics were (observed values in parenthesis): number of haplotypes in group 1 (4); number of segregating sites in group 1 (3); average number of pairwise differences in group 1 (0.5163); Tajima's *D* in group 1 (-1.131); number of haplotypes in group 2 (2); number of segregating sites in group 2 (1.0); average number of pairwise differences in

group 2 (0.1429); Tajima's *D* in group 2 (-1.155); number of haplotypes in group 3 (22); number of segregating sites in group 3 (20); average number of pairwise differences in group 3 (3.578); Tajima's *D* in group 3 (-1.104); average number of pairwise differences between groups 1 and 2 (0.3492); F_{ST} between groups 1 and 2 (0.121); 15) Average number of pairwise differences between groups 2 and 3 (2.802); F_{ST} between groups 2 and 3 (0.376).

Results

The relative tendencies showed by the different methods regarding the model likelihoods estimates were as follows: the logistic regression tended to produce higher estimates than the other methods when the model likelihoods were large (> 0.6) and smaller when they were small (< 0.1), while the direct approach and the kernel-beta method did not showed a sustained tendency respect each other. However, the model likelihoods estimated by the direct approach and the kernel-beta were more similar between them than with the logistic regression (Table 1). In fact, the differences was more than subtle; the average difference in the estimates was an order of magnitude smaller when the direct approach and the kernel-beta were compared than when the logistic regression was compared to both direct and kernel-beta.

When applied to the real dataset the estimations of the three methods also showed differences (Table 2), with the logistic regression picking a different model than the direct approach and the kernel-beta method.

Regarding the rates of error and the statistical power displayed by the three methods, they were, in general, similar but some clear tendencies were observed. The kernel-beta consistently showed the lowest rates of error with few exceptions, but also displayed a consistently lower statistical power than the other methods. However, the reduction in statistical power observed with the kernel-beta method was marginal and, in general, smaller than the loss in error rate (Table 3). On the other hand, the logistic regression presented an inconsistent behavior sometimes presenting the highest and the lowest statistical powers, as

well as the highest and the lowest error rates, sometimes in the same simulated example. Noticeably, the logistic regression presented a large number of failed estimations which increased with the complexity of the simulated examples. In fact, for simulated examples 4 and 5, the ones with the largest numbers of summary statistics, parameters and priors, 100% of the model choice procedures were aborted.

When investigated in detail, all the problems with logistic regression were numerical, i.e. caused by the particular ways that a computer approaches the analytic calculations by the discrete machine arithmetic. Those problems persisted even though regression was performed by the LAPACK package, which is a set of linear algebra routines developed for being error proof, and after all the numerical problems admitting direct solutions were fixed. The problems included: sensitiveness due to the s-shape of the logistic function, which produces large changes in the likelihoods estimations by small changes in the regression output; overflow; "catastrophic cancelation"; and linear dependency in the regression matrix, which was provoked by having repeated vectors of summary statistics in the reference table.

Discussion

The consistently lower rate of error in the kernel-beta method can be attributed primarily to the use of statistical error because, at the end, it was the only method incorporating error in the criteria for model choice since numerical issues prevented logistic regression from providing useable error estimates.

Direct approach had always higher statistical power and higher error rate, while logistic regression had a less predictable behavior. The lower error rate of the kernel-beta method was accompanied of marginal loses of statistical power, making the kernel-beta a conservative test.

Conservative analyses are less popular since they demise the conclusiveness of results. However, it is in the spirit of science in general to exert more control upon false positives (type II error) than upon false

negatives (type I error) under the argument that a false knowledge is more pernicious than no knowledge at all. A false knowledge could lead to dead ends provoking significant waste of resources and time at the least, while an inconclusive result serves to change or improve sampling, method or perspective, which would yield positive results in the middle term. In the specific case of the model choice by ABC, sometimes it is no quite clear which one is the null model, and all models could be considered equivalent, making meaningless the difference between type I and type II errors. In those cases the conservative behavior would be preserved by giving preference to error control rather than to statistical power. So, the kernel-beta method is indeed a conservative test.

However, the same behavior that could be considered advantageous could also be considered a drawback if statistical power is preferred due to a low error rate. In that case, it is reasonable to pursuit more statistical power, which could be afforded by parametric regression methods that have proved accurate in optimal circumstances (Beaumont 2002; Leuenberger & Wegmann 2010; Fagundes et al. 2007). Regression approaches, however, are not without problems.

Logistic regression as applied here suffered from a number of numerical issues. Some of them are treatable but some other could not only be unsolvable but also shared with other linear regression-based approaches. For example, any regression method employing least squares would fail if the regressors matrix is singular which occurs if two or more simulations get the same (analyzed) summary statistics. The chances of that occurring could be not so low with small sampling sizes, low mutation rates, and few summary statistics, in studies with molecular markers. In fact, that issue occurred in 60% of the simulations of the simulated example 1.

In addition, bias due to overparametrization and lack of assumptions fulfillment, the two problems that kernel-beta method is free from design, could also be notorious in regression methods. In the simulated examples 4 and 5 up to 100% of the simulations presented numerical issues. Those examples correspond

to the most complex simulated scenarios, where the number of parameters and summary statistics where the largest ones (14 and 21 summary statistics, and 4 and 8 parameters, respectively). In those examples the acceptance distances were large indeed, and so was the approaching error.

On the other hand, the kernel-beta method is the first method with a full Bayesian basis which could be advantageous for incorporating prior information and obtaining better error estimates but also its regression-type fundament makes it an interesting extension to the collection of regression methods that can be applied to ABC. In addition, its lack of parameters and assumptions could make it a standard reference, especially in high dimensional cases or when results among methods were too disparate, while parametric regression methods can be reserved to improve statistical power when error and numerical issues are under control (which can be assessed by PODs).

Despite the advantages of having a non-parametric and assumptions-free method, parametric regression still has the potential of greatly improve estimation when used under the right circumstances. So, a logical improvement could be to implement some techniques or criteria for assigning the pertinent type of regression that fits better an specific dataset, from a set of regression techniques that could include linear regression, generalized linear models (GLM), non-linear regression, non-parametric regression or regression by neural networks. Many of them have actually been already implemented for parameters estimation by ABC (e.g. Blum & Francois 2010).

Other lines of research should keep focusing in tackling the problems related with multidimensionality, as complexity of studies is always increasing, and the use of high throughput data (e.g. Next-Generation Sequencing data). Since, high throughput data has been called to dominate molecular studies in ecology and evolution, it will be important to keep an eye in the computer efficiency of new developments as well.

Acknowledgements

I thank Love Dalén for support and feedback. This work was supported by Formas via the ERA-NET Biodiversa project Climigrate and the Strategic Research Programme Ekoklim at Stockholm University.

References

Beaumont MA (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annu Rev Ecol Evol S* **41,** 379-406.

Beaumont MA, Rannala B (2004) The Bayesian revolution in genetics. *Nat Rev Genet* **5**(4), 251-261.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* **162**(4), 2025-2035.

Bertorelle G, Benazzo A, Mona S (2010) ABC as a flexible framework to estimate demography over space and time: some cons, many pros. *Mol Ecol* **19**(13), 2609-2625.

Blum MGB, Francois O (2010) Non-linear regression models for Approximate Bayesian Computation. *Stat Comput* **20**(1), 63-73.

Csillery K, Blum MGB, Gaggiotti OE, Francois O (2010) Approximate Bayesian Computation (ABC) in practice. *Trends Ecol Evol* **25**(7), 410-418.

Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano FM, et al. (2007) Statistical evaluation of alternative models of human evolution. *P Natl Acad Sci USA* **104**(45), 17614-17619.

Gelman A, Carlin JB, Stern HS, Rubin DB (2003). *Bayesian Data Analysis*. Chapman and Hall, London, UK.

Hamilton G, Currat M, Ray N, Heckel G, Beaumont M, et al. (2005) Bayesian estimation of recent migration rates after a spatial expansion. *Genetics* **170**(1), 409-417.

Hoban S, Bertorelle G, Gaggiotti OE (2012) Computer simulations: tools for population and evolutionary genetics. *Nat Rev Genet* **13**, 110-122.

Ilves L, Huang W, Wares JP and Hickerson MJ (2010) Colonization and/or mitochondrial selective sweeps across the North intertidal assemblage revealed by multi-taxa approximate Bayesian computation. *Mol Ecol* **19**(20), 4505-4519.

Jensen JD, Thornton KR, Andolfatto P (2008) An Approximate Bayesian Estimator Suggests Strong, Recurrent Selective Sweeps in Drosophila. *PLoS Genetics* **4**(9), e1000198.

Lagerholm VK, Sandoval-Castellanos E, Ehrich D, Abramson NI, Nadachowski A, et al. (2014) On the origin of the Norwegian lemming. *Mol Ecol* **23**(8), 2060-2071.

Leuenberger C, Wegmann D (2010) Bayesian Computation and Model Selection Without Likelihoods. *Genetics* **184**(1), 243-252.

Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. *P Natl Acad Sci USA* **100**(26), 15324-15328.

Nystrom V, Dalén L, Vartanyan S, Lidén K, Ryman N, et al. (2010) Temporal genetic change in the last remaining population of woolly mammoth. *Proc R Soc B* **277**(1692), 2331-2337.

Palkopoulou E, Dalén L, Lister AM, Vartanyan S, Sablin M, et al. (2013) Holarctic genetic structure and range dynamics in the woolly mammoth. *Proc R Soc B* **280**(1770), 20131910.

Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Mol Biol Evol* **16**(12): 1791-1798.

Ratmann O, Andrieu C, Wiuf C, Richardson S (2009) Model criticism based on likelihood-free inference, with an application to protein network evolution. *P Natl Acad Sci USA* **106**(26), 10576-10581.

Robert CP, Cornuet JM, Marin JM, Pillai NS (2011) Lack of confidence in approximate Bayesian computation model choice. *P Natl Acad Sci USA* **108**(37), 15112-15117.

Roux C, Tsagkogeorga G, Bierne N, Galtier N (2013) Crossing the Species Barrier: genomic Hotspots of Introgression between Two Highly Divergent *Ciona intestinalis* Species. *Mol Biol Evol* **30**(7), 1574-1587.

Sandoval-Castellanos E, Palkopoulou E, Dalén L (2014) Back to BaySICS: A User-Friendly Program for Bayesian Statistical Inference from Coalescent Simulations. *PLoS One* **9**(5), e98011.

Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. *P Natl Acad Sci USA* **104**(6), 1760-1765.

Sousa VC, Fritz M, Beaumont MA, Chikhi L (2009) Approximate Bayesian computation without summary statistics: the case of admixture. *Genetics* **181**(4), 1507-1519.

Sjodin P, Sjostrand AE, Jakobsson M, Blum MGB (2012) Resequencing Data Provide No Evidence for a Human Bottleneck in Africa during the Penultimate Glacial Period. *Mol Biol Evol* **29**(7), 1851-1860.

Tanaka MM, Francis AR, Luciani F, Sisson SA (2006) Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. *Genetics* **173**(3), 1511-1520.

Warmuth V, Eriksson A, Bower MA, Barker G, Barrett E et al. (2012) Reconstructing the origin and spread of horse domestication in Eurasian steppe. *P Natl Acad Sci USA* **109**(21), 8202-8206.

Wegmann D, Leuenberger C, Excoffier L (2009) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182**(4), 1207-1218.

Data Accesibility

The beta estimation is intended to constitute only a part of an ABC procedure, which full set of instructions could be sophisticated and large. So the provided source code, where the instructions for a beta estimation occupy less than 20 lines, is intended rather as a guide for implementation in the platform of one's predilection. Source code and a windows executable are available at: https://www.dropbox.com/sh/11pe1obb8j46qf4/AABuUoekq-QkRN3EspTGqzena. In addition the kernel-beta method is implemented in the software BaySICS (Sandoval-Castellanos et al. 2014).

Figure Legends

Figure 1. Schematic representation of the scenarios (models) tested in the five simulated examples and the real dataset. A-C represent the three scenarios of the simulated example 1. D-F represent the scenarios tested in the simulated example 4, which had six scenarios: the three displayed without migration between the populations and the same three with migration among the populations. G represent the two scenarios tested in the simulated example 5; the difference between the two scenarios concerns only the range of T_1: 10,000-15,000 years BP for one scenario and 30,000-120,000 for the other one. H-K represents the scenarios tested with the real data, the ages of the changes (in calendar years) at right. H-K also represent the scenarios of the simulated examples 2 and 3 with two differences: the simulated examples had only three stages (instead the four shown) but keeping the same structure; and the times were different, being the youngest time the present time (instead 3,000 as displayed) and the two changes occurring at 5,000 and 10,000 years before present (BP) (instead 9,000 and 12,000). The simulated examples 2 and 3 were only different in the ages of the samples: in the simulated example 2 all the samples were contemporary while in the simulated example 3 all the samples had ages between 0.0 and 19,755 years BP.

Table 1. Average differences among model likelihoods estimates. The differences were averaged among every iteration and also among the different scenarios tested in each simulated example.

	Direct vs Kernel-Beta	**Direct vs Logistic reg.**	**Kernel-beta vs Logistic reg.**
Simulated example 1	0.0210	0.1203	0.1209
Simulated example 2	0.0053	0.1474	0.1466
Simulated example 3	0.0087	0.0715	0.0647
Simulated example 4	0.0101	0.1666	0.1667
Simulated example 5	0.0136	0.0688	0.0775

Table 2 Model likelihoods for the mammoth dataset as obtained with the three methods. Numbers in bold correspond to the scenarios that were chosen by the respective method.

Competing models	Direct	Logistic	Beta estimation
Null	0.0105	0.0048	0.0065
Reduction	0.0124	0.0003	0.0073
Bottleneck	**0.6892**	0.3727	**0.7125**
Two Sizes	0.2880	**0.5503**	0.2735

509 Table 3. Statistical power and type I/II error rate. Statistical power is the naked number while error
510 appears in square brackets. Rates were obtained with the three different methods: the direct approach; the
511 logistic regression in original way (number of right and wrong choices divided by the total number
512 simulations) and adjusted (number of right and wrong choices divided by the number of simulations that
513 succeeded in the estimation –i.e. without numerical errors-); and the kernel-beta method (always with
514 statistical error).

515

516

517

518

519

520

521

522

523

524

525

526

527

Competing models	Direct	Logistic regression		Kernel-beta
		adjusted	original	
Example1				
Null	10.6 [89.4]	10.5 [89.5]	4.1 [35.0]	6.0 [82.6]
Expansion	64.2 [35.8]	48.9 [51.2]	12.5 [13.1]	61.5 [31.2]
Reduction	55.6 [44.4]	70.1 [29.0]	31.3 [12.8]	52.1 [37.8]
Example 2				
Null	13.3 [86.7]	20.0 [80.0]	8.7 [34.8]	11.6 [80.4]
Reduction	76.2 [23.8]	87.2 [12.7]	46.4 [6.8]	75.2 [22.1]
Bottleneck	79.2 [21.0]	63.5 [36.5]	20.5 [11.8]	77.8 [18.5]
Two Sizes	29.1 [70.9]	29.5 [70.5]	11.9 [28.5]	24.0 [65.9]
Example 3				
Null	42.0 [58.0]	47.7 [52.3]	46.5 [50.9]	35.3 [42.0]
Reduction	45.0 [55.0]	49.4 [50.6]	48.6 [49.7]	38.1 [42.4]
Bottleneck	81.5 [18.5]	90.8 [9.2]	88.4 [9.0]	81.7 [11.8]
Two Sizes	54.3 [45.7]	40.8 [59.2]	29.4 [42.7]	49.1 [39.9]
Example 4				
Equal	12.4 [87.6]	N/A	N/A	13.2 [73.4]
Pop 1 > Pop 2	40.6 [59.43]	N/A	N/A	39.1 [51.0]
Pop 1 < Pop 2	42.4 [57.6]	N/A	N/A	40.4 [48.5]
Equal + mig	23.8 [76.2]	N/A	N/A	22.6 [63.6]
Pop 1 > Pop 2 + mig	37.6 [62.4]	N/A	N/A	33.5 [53.9]
Pop 1 < Pop 2 + mig	40.0 [60.0]	N/A	N/A	35.2 [51.2]
Example 5				
Younger split	45.6 [54.4]	N/A	N/A	44.1[39.4]
Older split	83.7 [16.3]	N/A	N/A	74.0 [14.6]

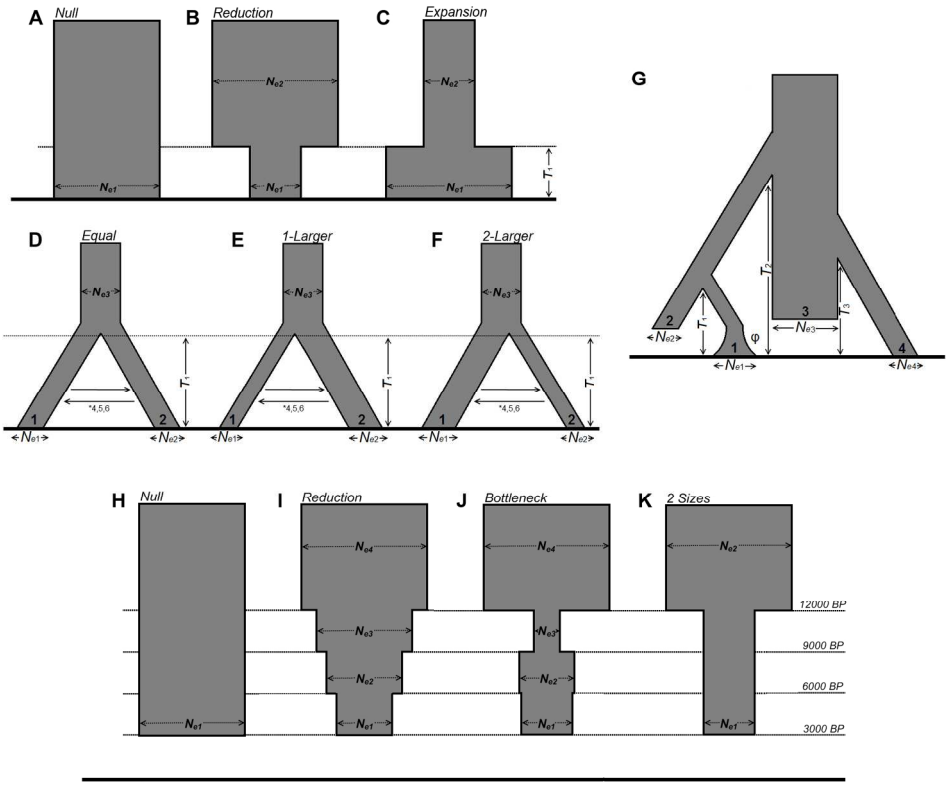

Figure 1. Schematic representation of the scenarios (models) tested in the five simulated examples and the real dataset. A-C represent the three scenarios of the simulated example 1. D-F represent the scenarios tested in the simulated example 4, which had six scenarios: the three displayed without migration between the populations and the same three with migration among the populations. G represent the two scenarios tested in the simulated example 5; the difference between the two scenarios concerns only the range of T1: 10,000-15,000 years BP for one scenario and 30,000-120,000 for the other one. H-K represents the scenarios tested with the real data, the ages of the changes (in calendar years) at right. H-K also represent the scenarios of the simulated examples 2 and 3 with two differences: the simulated examples had only three stages (instead the four shown) but keeping the same structure; and the times were different, being the youngest time the present time (instead 3,000 as displayed) and the two changes occurring at 5,000 and 10,000 years before present (BP) (instead 9,000 and 12,000). The simulated examples 2 and 3 were only different in the ages of the samples: in the simulated example 2 all the samples were contemporary while in the simulated example 3 all the samples had ages between 0.0 and 19,755 years BP.
529x441mm (96 x 96 DPI)