Regression approaches for Approximate Bayesian Computation

Michael G B Blum

Michael G.B. Blum, CNRS Université Grenoble Alpes

Regression approaches for Approximate Bayesian Computation

0.1 Content

This chapter introduces regression approaches and regression adjustment for Approximate Bayesian Computation (ABC). Regression adjustment adjusts parameter values after rejection sampling in order to account for the imperfect match between simulations and observations. Imperfect match between simulations and observations can be more pronounced when there are many summary statistics, a phenomenon coined as the *curse of dimensionality* [5]. Because of this imperfect match, credibility intervals obtained with regression approaches can be inflated compared to true credibility intervals [10]. The chapter presents the main concepts underlying regression adjustment. A theorem that compares theoretical properties of posterior distributions obtained with and without regression adjustment is presented. Last, a practical application of regression adjustment in population genetics shows that regression adjustment shrinks posterior distributions compared to rejection approaches, which is a solution to avoid inflated credibility intervals.

0.2 Introduction

In this chapter, we present regression approaches for Approximate Bayesian Computation (ABC). As for most methodological developments related to ABC, regression approaches originate with coalescent modeling in population genetics [3]. After performing rejection sampling by accepting parameters that generate summary statistics close enough to those observed, parameters are *adjusted* to account for the discrepancy between simulated and observed summary statistics. Because adjustment is based on a regression model, such approaches are coined as *regression adjustment* in the following.

Regression adjustment is a peculiar approach in the landscape of Bayesian approaches where sampling techniques are usually proposed to account for mismatches between simulations and observations [18, 30]. We suggest various reasons explaining why regression adjustment is now a common step in practical applications of ABC. First, it is convenient and generic because the simulation mechanism is used to generate simulated summary statistics as a first step and it is not used afterwards. For instance, the software *ms* is used to generate DNA sequences or genotypes when performing ABC inference in population genetics [15]. Statistical routines, which account for mismatches, are completely separated from the simulation mechanism and are used in a second step. Regression adjustment can therefore be readily applied in a wide range of contexts without implementation efforts. By contrast, when considering sampling techniques, statistical operations and simulations are embedded within a single algorithm [30, 2], which may require new algorithmic devel-

opment for each specific statistical problem. Second, regression approaches have been shown to produce reduced statistical errors compared to rejection algorithms in a quite diverse range of statistical problems [3, 6, 27]. Last, regression approaches are implemented in different ABC software including *DIYABC* [9] and the R *abc* package [11].

In this chapter, I introduce regression adjustment using a comprehensive framework that includes linear adjustment [3] as well as more flexible adjustments such as non-linear models [6]. The first section presents the main concepts underlying regression adjustment. The second section presents a theorem that compares theoretical properties of posterior distributions obtained with and without regression adjustment. The third section presents a practical application of regression adjustment in ABC. It shows that regression adjustment shrinks posterior distributions when compared to a standard rejection approach. The fourth section presents recent regression approaches for ABC that are not based on regression adjustment.

0.3 Principle of regression adjustment

0.3.1 Partial posterior distribution

Bayesian inference is based on the posterior distribution defined as

$$\pi(\theta|y_{\text{obs}}) \propto p(y_{\text{obs}}|\theta)\pi(\theta) \tag{1}$$

where $\theta \in \mathbb{R}^p$ is the vector of parameters, and y_{obs} are the data. Up to a renormalizing constant, the posterior distribution depends on the prior $\pi(\theta)$ and on the likelihood function $p(y_{\text{obs}}|\theta)$. In the context of ABC, inference is no longer based on the posterior distribution $\pi(\theta|y_{\text{obs}})$ but on the *partial* posterior distribution $\pi(\theta|s_{\text{obs}})$ where s_{obs} is a q-dimensional vector of descriptive statistics. The partial posterior distribution is defined as follows

$$\pi(\theta|s_{\text{obs}}) \propto p(s_{\text{obs}}|\theta)\pi(\theta). \tag{2}$$

Obviously, the partial posterior is equal to the posterior if the descriptive statistics s_{obs} are sufficient for the parameter θ.

0.3.2 Rejection algorithm followed by adjustment

To simulate a sample from the partial posterior $p(\theta|s_{obs})$, the rejection algorithm followed by adjustment works as follows

1. Simulate n values $\theta^{(i)}$, $i = 1, \ldots, n$, according to the prior distribution π.
2. Simulate descriptive statistics $s^{(i)}$ using the generative model $p(s^{(i)}|\theta^{(i)})$.
3. Associate with each pair $(\theta^{(i)}, s^{(i)})$ a weight $w^{(i)} \propto K_h(\|s^{(i)} - s_{\text{obs}}\|)$ where $\|\cdot - \cdot\|$ is a distance function, $h > 0$ is the bandwidth parameter, and K is an univariate statistical kernel with $K_h(\|\cdot\|) = K(\|\cdot\|/h)$.
4. Fit a regression model where the response is θ and the predictive variables are the summary statistics s (equations (3) or (5)). Use a regression model to adjust the $\theta^{(i)}$ in order to produce a weighted sample of adjusted values. Homoscedastic adjustment (equation (4)) or heteroscedastic adjustment (equation (6)) can be used to produce a weighted sample $(\theta_c^{(i)}, w^{(i)})$, $i = 1, \ldots, n$, which approximates the posterior distribution.

To run the rejection algorithm followed by adjustment, there are several choices to make. The first choice concerns the kernel K. Usual choices for K encompass uniform kernels that give a weight of 1 to all accepted simulations and zero otherwise [23] or the Epanechnikov kernel for a smoother version of the rejection algorithm [3]. However, as for traditional density estimation, the choice of statistical kernel has a weak impact on estimated distribution [29]. The second choice concerns the threshold parameter h. For kernels with a finite support, the threshold h corresponds to (half) the window size within which simulations are accepted. For the theorem presented in section 0.5, I assume that h is chosen without taking into account the simulations $s^{(1)}, \ldots, s^{(n)}$. This technical assumption does not hold in practice where we generally choose to accept a given percentage p, typically 1% or 0.1%, of the simulations. This practice amounts at setting h to the first p-quantile of the distances $\|s^{(i)} - s_{\text{obs}}\|$. A theorem where the threshold depends on simulations has been provided [4]. Choice of threshold h corresponds to bias-variance tradeoff. When choosing small values of h, the number of accepted simulations is small and estimators might have a large variance. By contrast, when choosing large values of h, the number of accepted simulations is large and estimators might be biased [5].

0.3.3 Regression adjustment

The principle of regression adjustment is to adjust simulated parameters $\theta^{(i)}$ with nonzero weights $w^{(i)} > 0$ in order to account for the difference between

the simulated statistics $s^{(i)}$ and the observed one s_{obs}. To adjust parameter values, a regression model is fitted in the neighborhood of s_{obs}

$$\theta^{(i)} = m(s^{(i)}) + \varepsilon, \quad i = 1, \cdots, n \tag{3}$$

where $m(s)$ is the conditional expectation of θ given s and ε is the residual. The regression model of equation (3) assumes *homoscedasticity*, i.e. it assumes that the variance of the residuals does not depend on s. To produce samples from the partial posterior distribution, the $\theta^{(i)}$'s are adjusted as follows

$$\begin{aligned} \theta_c^{(i)} &= \hat{m}(s_{\text{obs}}) + \hat{\varepsilon}^{(i)} \\ &= \hat{m}(s_{\text{obs}}) + (\theta^{(i)} - \hat{m}(s^{(i)})), \end{aligned} \tag{4}$$

where $\hat{m}$ represents an estimator of the conditional expectation of θ given s, and $\hat{\varepsilon}^{(i)}$ is the i^{th} empirical residual. In its original formulation, regression adjustment assumes that m is a linear function [3] and it was later extended to non-linear adjustments [6].

The homoscedastic assumption of equation (3) may not be always valid. When the number of simulations is not very large because of computational constraints, local approximations, such as the homoscedastic assumption, are no longer valid because the neighborhood corresponding to simulations for which $w^{(i)} \neq 0$ is too large. Regression adjustment can account for heteroscedasticity that occurs when the variance of the residuals depend on the summary statistics. When accounting for heteroscedasticity, the regression equation can be written as follows [6]

$$\theta^{(i)} = m(s^{(i)}) + \sigma(s^{(i)})\zeta, \quad i = 1, \cdots, n \tag{5}$$

where $\sigma(\mathbf{s})$ is the square root of the conditional variance of θ given s, and ζ is the residual. Heteroscedastic adjustment involves an additional scaling step in addition to homoscedastic adjustment (4) (Figure 1)

$$\begin{aligned} \theta_{c'}^{(i)} &= \hat{m}(s_{\text{obs}}) + \hat{\sigma}(s_{\text{obs}})\hat{\zeta}^{(i)} \\ &= \hat{m}(s_{\text{obs}}) + \frac{\hat{\sigma}(s_{\text{obs}})}{\hat{\sigma}(s^{(i)})}(\theta^{(i)} - \hat{m}(s^{(i)})), \end{aligned} \tag{6}$$

where $\hat{m}$ and $\hat{\sigma}$ are estimators of the conditional mean and of the conditional standard deviation.

0.3.4 Fitting regression models

Equations (4) and (6) of regression adjustment depend on the estimator of the conditional mean $\hat{m}$ and possibly of the conditional variance $\hat{\sigma}$. Model fitting is performed using weighted least squares. The conditional mean is learned by minimizing the following weighted least square criterion

$$E(m) = \sum_{i=1}^{n} (\theta^{(i)} - m(s^{(i)}))^2 w^{(i)}. \tag{7}$$

For linear adjustment, we assume that $m(s) = \alpha + \beta s$ [3]. The parameters α and β are inferred by minimizing the weighted least square criterion given in equation (7).

For heteroscedastic adjustment (equation (4)), the conditional variance should also be inferred. The conditional variance is learned after minimization of a least square criterion. It is obtained by fitting a regression model where the answer is the logarithm of the squared residuals. The weighted least squares criterion is given as follows

$$E(\log \sigma^2) = \sum_{i=1}^{n} \left(\log((\hat{\varepsilon}^{(i)})^2) - \log \sigma^2(s) \right)^2 w^{(i)}.$$

Neural networks has been proposed to estimate m and σ^2 [6]. This choice was motivated by the possibility offered by neural networks to reduce the dimension of descriptive statistics via an internal projection on a space of lower dimension [25].

In general, the assumptions of homoscedasticity and linearity (equation (3)) are violated when the percentage of accepted simulation is large. By contrast, heteroscedastic and non-linear regression models (equation (5)) are more flexible. Because of this additional flexibility, the estimated posterior distributions obtained after heteroscedastic and non-linear adjustment is less sensitive to the percentage of accepted simulations [6]. In a coalescent model where the objective was to estimate the mutation rate, heteroscedastic adjustment with neural networks was found to be less sensitive to the percentage of accepted simulations than linear and homoscedastic adjustment [6]. In a model of phylodynamics, it was found again that statistical error obtained with neural networks decreases at first—because the regression method requires a large enough training dataset—and then reaches a plateau [27]. However for larger phylodynamics dataset, statistical error obtained with neural networks increases for higher tolerance values. Poor regularization or the limited size of neural networks were advanced as putative explanations [27].

In principle, estimation of the conditional mean m and of the conditional variance σ^2 can be performed with different regression approaches. For instance, the `R` *abc* package implements different regression models for regression adjustment including linear regression, ridge regression and neural networks [11]. Lasso regression is another regression approach that can be considered. Regression adjustment based on lasso was shown to provide smaller errors than neural network in a phylodynamic model [27]. An advantage of lasso, ridge regression and neural networks compared to standard multiple regression is that they account for the large dimension of the summary statistics using different regularization techniques. Instead of considering regularized regression, dimension reduction is an alternative where the initial summary statistics are replaced by a reduced set of summary statistics or a combination of the initial summary statistics [12, 7]. The key and practical advantage of regression approaches with regularization is that they implicitly account for

the large number of summary statistics and the additional step of variable selection can be avoided.

0.3.5 Parameter transformations

When the parameters are bounded or positive, parameters can be transformed before regression adjustment. Transformations guarantee that the adjusted parameter values lie in the range of the prior distribution [3]. An additional advantage of the *log* and *logit* transformations is that they stabilize the variance of the regression model and make regression model (3) more homoscedastic [5].

Positive parameters are regressed on a logarithm scale $\phi = \log(\theta)$,

$$\phi = m(\mathbf{s}) + \varepsilon.$$

Parameters are then adjusted on the logarithm scale

$$\phi_c^{(i)} = \hat{m}(s_{\text{obs}}) + (\phi^{(i)} - \hat{m}(s^{(i)})).$$

The final adjusted values are obtained by exponentiation of the adjusted parameter values

$$\theta_c^{(i)} = \exp(\phi_c^{(i)}).$$

Instead of using a logarithm transformation, bounded parameters are adjusted using a logit transformation. Heteroscedastic adjustment can also be performed after log or logit transformations.

0.3.6 Shrinkage

An important property of regression adjustment concerns posterior shrinkage. When considering linear regression, the empirical variance of the residuals is smaller than the total variance. In addition, residuals are centered for linear regression. These two properties imply that for linear adjustment, the empirical variance of $\theta_c^{(i)}$ is smaller than the empirical variance of the non-adjusted values $\theta^{(i)}$ obtained with the rejection algorithm. Following homoscedastic and linear adjustment, the posterior variance is consequently reduced. For non-linear adjustment, shrinkage property has also been reported and the additional step generated by heteroscedastic adjustment does not necessarily involve additional shrinkage when comparing $\theta_c^{(i)}$ to $\theta_{c'}^{(i)}$ [5].

0.4 Theoretical results about regression adjustment

The following theoretical section is technical and can be skipped by readers not interested by mathematical results about ABC estimators based on regression adjustment. In this section, we give the main theorem that describes

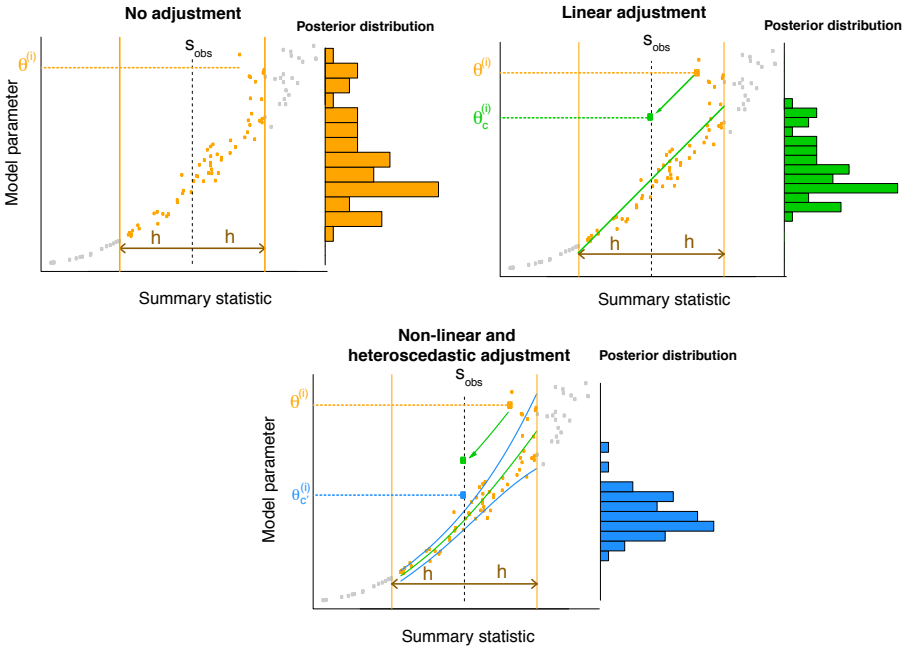

FIGURE 1
Posterior distributions obtained with and without regression adjustment. The shrinking effect of adjustment is visible since the posterior variance of the green and blue histograms is reduced compared to the posterior variance of the orange histogram.

the statistical properties of posterior distributions obtained with or without regression adjustment. To this end, the estimators of the posterior distribution are defined as follows

$$\hat{\pi}_j(\theta|s_{\text{obs}}) = \sum_{i=1}^{n} \tilde{K}_{h'}(\theta_j^{(i)} - \theta) w^{(i)},\ j = 0, 1, 2, \tag{8}$$

where $\theta_0^{(i)} = \theta^{(i)}$ (no adjustment), $\theta_j^{(i)} = \theta_c^{(i)}$ for $j = 1, 2$ (homoscedastic adjustment), $\tilde{K}$ is an univariate kernel, and $\tilde{K}_{h'}(\cdot) = \tilde{K}(\cdot)/h'$. Linear adjustment corresponds to $j = 1$ and quadratic adjustment corresponds to $j = 2$. In nonparametric statistics, estimators of the conditional density with adjustment have already been proposed [16, 13].

To present the main theorem, we introduce the following notations: if X_n is a sequence of random variables and a_n is a deterministic sequence, the notation $X_n = o_P(a_n)$ means that X_n/a_n converges to zero in probability and $X_n = O_P(a_n)$ means that the ratio X_n/a_n is bounded in probability when n goes to infinity. The technical assumptions of the theorem are given in the appendix of [5].

Theorem 1 *We assume that conditions (A1)-(A5) given in the appendix of [5] hold. The bias and variance of the estimators $\hat{\pi}_j(\theta|s_{\text{obs}})$, $j = 0, 1, 2$ are given by*

$$E[\hat{\pi}_j(\theta|s_{\text{obs}}) - \pi(\theta|s_{\text{obs}})] = C_1 h'^2 + C_{2,j} h^2 + O_P((h^2 + h'^2)^2) + O_P(\frac{1}{nh^q}), \tag{9}$$

$$\text{Var}[\hat{\pi}_j(\theta|s_{\text{obs}})] = \frac{C_3}{nh^q h'}(1 + o_P(1)), \tag{10}$$

where q is the dimension of the vector of summary statistics and the constants C_1, $C_{2,j}$ et C_3 are given in [5].

Proof: See [5].

There are other theorems that provide asymptotic biases and variances of ABC estimators but they do not study the properties of estimators arising after regression adjustment. Considering posterior expectation (e.g. posterior moments) instead of the posterior density, [1] provides asymptotic bias and variance of an estimator obtained with rejection algorithm. [4] studied asymptotic properties when the window size h depends on the data instead of being fixed in advance. Another version of ABC called *lazy ABC* exists and provides a bias proportional to h instead of h^2 while keeping the same variance of the order of $1/(nh^q)$, which is inversely proportional to the acceptance probability in the rejection algorithm [12].

Remark 1. Curse of dimensionality The mean square error of the estimators is the sum of the squared bias and of the variance. With elementary algebra, we can show that for the three estimators $\hat{\pi}_j(\theta|s_{\text{obs}})$, $j = 0, 1, 2$, the mean square error is of the order of $n^{-1/(q+5)}$ for an optimal choice of h. The

speed with which the error approaches 0 therefore decreases drastically when the dimension of the descriptive statistics increases. This theorem highlights (in an admittedly complicated manner) the importance of reducing the dimension of the statistics. However, the findings from these asymptotic theorems, which are classic in non-parametric statistics, are often much more pessimistic than the results observed in practice. It is especially true because asymptotic theorems in the vein of Theorem 1 do not take into account correlations between summary statistics [28].

Remark 2. Comparing biases of estimators with and without adjustment It is not possible to compare biases (i.e. the constant $C_{2,j}$, $j = 0, 1, 2$) for any statistical model. However, if we assume that the residual distribution of ε in equation (3) does not depend on s, then the constant $C_{2,2}$ is 0. When assuming homoscedasticity, the estimator that achieves asymptotically the smallest mean square error is the estimator with quadratic adjustment $\hat{p}_2(\theta|s_{\text{obs}})$. Assuming additionally that the conditional expectation m is linear in s, then both $\hat{p}_1(\theta|s_{\text{obs}})$ and $\hat{p}_2(\theta|s_{\text{obs}})$ have a mean square error lower than the error obtained without adjustment.

0.5 Application of regression adjustment to estimate admixture proportions using polymorphism data

To illustrate regression adjustment, I consider an example of parameter inference in population genetics. Description of coalescent modeling in population genetics is out of the scope of this chapter and we refer interested readers to dedicated reviews [14, 26]. This example illustrates that ABC can be used to infer evolutionary events such as admixture between sister species. I assume that two populations (A and B) diverged in the past and admixed with admixture proportions p and $1 - p$ to form a new hybrid species C that subsequently split to form two sister species C_1 and C_2 (Figure 2). Simulations are performed using the software DIYABC [9]. The model of Figure 2 corresponds to a model of divergence and admixture between species of a complex of species from the butterfly gender *Coenonympha*. We assume that 2 populations of the Darwin's Heath (*Coenonympha darwiniana*) originated through hybridization between the Pearly Heath (*Coenonympha arcania*) and the Alpine Heath (*Coenonympha gardetta*) [8]. A total of 16 summary statistics based on Single Nucleotide Polymorphisms (SNPs) are used for parameter inference [8]. A total of 10^6 simulations are performed and the percentage of accepted simulations is of 0.5%.

I consider four different forms of regression adjustment: linear and homoscedastic adjustment, non-linear (neural networks) and homoscedastic adjustment, linear and heteroscedastic adjustment, non-linear and heteroscedastic adjustment. All adjustments were performed with the `R` package *abc*

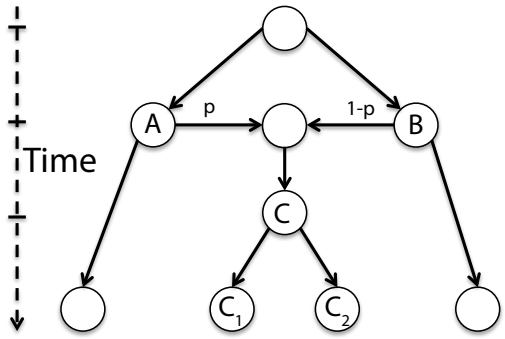

FIGURE 2
Graphical description of the model of admixture between sister species. Two populations (A and B) diverge in the past and admixed with admixture proportions p and $1-p$ to form a new hybrid species C that subsequently diverge to form two sister species (C_1 and C_2).

[11, 24]. I evaluate parameter inference using a cross-validation criterion [11]. The cross-validation error decreases when considering linear adjustment (Figure 3). However, considering heteroscedastic instead of homoscedastic adjustment does not provide an additional decrease of the cross-validation error (Figure 3).

Then, using real data from a butterfly species complex, we compare the posterior distribution of the admixture proportion p obtained without adjustment, with linear and homoscedastic adjustment, and with non-linear and homoscedastic adjustment (Figure 4). For this example, considering regression adjustment considerably changes the shape of the posterior distribution. The posterior mean without adjustment is of $p = 0.51$ ($95\%\, C.I. = (0.12, 0.88)$). By contrast, when considering linear and homoscedastic adjustment, the posterior mean is of 0.93 ($95\%\, C.I. = (0.86, 0.98)$. When considering non-linear and homoscedastic adjustment, the posterior mean is 0.84 ($95\%\, C.I. = (0.69, 0.93)$. Regression adjustment confirms a larger contribution of *C. arcania* to the genetic composition of the ancestral *C. darwiniana* population [8]. This example shows that regression adjustment not only shrinks credibility intervals but can also shift posterior estimates. Compared to rejection, the posterior shift observed with regression adjustments provides a result that is more consistent with published results [8].

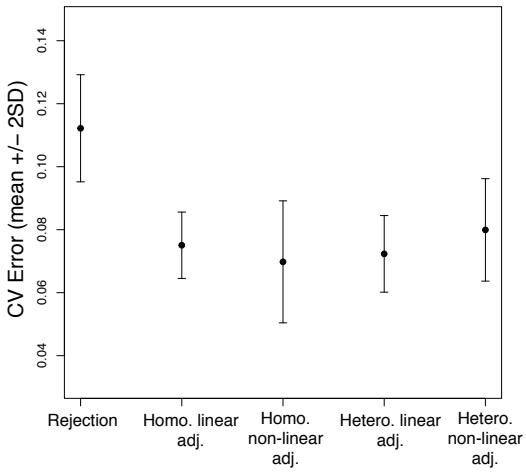

FIGURE 3
Errors obtained when estimating the admixture proportion p with different ABC estimators. The errors are obtained with cross-validation and error bars (two standard deviations) are estimated with bootstrap. *adj.* stands for adjustment.

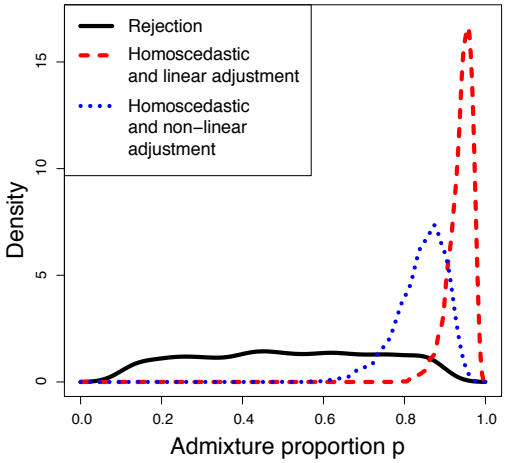

FIGURE 4
Posterior distribution of the admixture coefficient p obtained using real data from a butterfly species complex to compute observed summary statistics. This example shows that regression adjustment not only shrinks credibility intervals but can also shift posterior estimates. The admixture coefficient p measures the relative contribution of *Coenonympha arcania* to the ancestral *C. darwiniana* population.

0.6 Regression methods besides regression adjustment

There are other regression methods besides regression adjustment that have been proposed to estimate $E[\theta|s_{obs}]$ and $\pi(\theta|s_{obs})$ using the simulations as a training set. A first set of methods consider kernel methods to perform regression [21]. The principle is to define a kernel function K to compare observed statistics to simulated summary statistics. Because of the so-called kernel trick, regression with kernel methods amounts at regressing θ with $\Phi(s)$ where $< \Phi(s), \Phi(s') >= K(s, s')$ for two vector of summary statistics s and s'. Then, an estimate of the posterior mean is obtained as follows

$$E[\theta|s_{obs}] = \sum_{i=1}^{n} w_i \theta^{(i)}, \tag{11}$$

where w_i depends on the inverse the Gram matrix containing the values $K(s^{(i)}, s^{(j)})$ for $i, j = 1 \ldots, n$. A formula to estimate posterior density can also be obtained in the same lines as formula (11). Simulations suggest that kernel ABC gives better performance than regression adjustment when high-dimensional summary statistics are used. For a given statistical error, it was reported that fewer simulations should be performed when using kernel ABC instead of regression adjustment [21]. Other kernel approaches have been proposed for ABC where simulated and observed samples or summary statistics are directly compared through a distance measure between empirical probability distributions [20, 22].

Another regression method in ABC that does not use regression adjustment considers quantile regression forest [17]. Generally used to estimate conditional mean, random forests also provide information about the full conditional distribution of the response variable [19]. By inverting the estimated conditional cumulative distribution function of the response variable, quantiles can be inferred [19]. The principle of quantile regression forest is to use random forests in order to give a weight w_i to each simulation $(\theta^{(i)}, s^{(i)})$. These weights are then used to estimate the conditional cumulative posterior distribution function $F(\theta|s_{obs})$ and to provide posterior quantiles by inversion. An advantage of quantile regression forest is that tolerance rate should not be specified and standard parameters of random forest can be considered instead. A simulation study of coalescent models shows that regression adjustment can shrink posterior excessively by contrast to quantile regression forest [17].

0.7 Conclusion

This chapter introduces regression adjustment for Approximate Bayesian Computation [3, 6]. We explain why regression adjustment shrinks posterior distribution which is a desirable feature because credibility intervals obtained with rejection methods can be too wide [5]. When inferring admixture with SNP data in a complex of butterfly species, the posterior distribution obtained with regression adjustment was not only shrunk when compared to standard rejection but also shifted to larger values, which confirm results obtained for this species complex with other statistical approaches [8]. We have introduced different variants of regression adjustment and it might be difficult for ABC users to choose which adjustment is appropriate in their context. We argue that there is no best strategy in general. In the admixture example, we found, based on a cross-validation error criterion, that homoscedastic linear adjustment provides considerable improvement compared to rejection. More advanced adjustments provide almost negligible improvements if no improvement at all. However, in a model of phylodynamics, non-linear adjustment was reported to achieve considerable improvement compared to linear adjustment [27]. In practical applications of ABC, we suggest to compute errors such as cross-validation estimation errors to choose a particular method for regression adjustment.

With the rapid development of complex machine learning approaches, we envision that regression approaches for Approximate Bayesian Computation can be further improved to provide more reliable inference for complex models in biology and ecology.

Bibliography

[1] Stuart Barber, Jochen Voss, and Mark Webster. The rate of convergence for Approximate Bayesian Computation. *Electronic Journal of Statistics Electronic Journal of Statistics Electronic Journal of Statistics*, 9:80–105, 2015.

[2] Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert. Adaptive approximate Bayesian computation. *Biometrika*, 96(4):983–990, 2009.

[3] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

[4] Gérard Biau, Frédéric Cérou, Arnaud Guyader, et al. New insights into approximate Bayesian computation. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 51, pages 376–403. Institut Henri Poincaré, 2015.

[5] Michael G. B. Blum. Approximate Bayesian Computation: A nonparametric perspective. *Journal of the American Statistical Association*, 105(491):1178–1187, 2010.

[6] Michael G B Blum and Olivier François. Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, 20:63–73, 2010.

[7] Michael GB Blum, David Nunes, Dennis Prangle, Scott A Sisson, et al. A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, 28(2):189–208, 2013.

[8] Thibaut Capblancq, Laurence Després, Delphine Rioux, and Jesús Mavárez. Hybridization promotes speciation in coenonympha butterflies. *Molecular ecology*, 24(24):6209–6222, 2015.

[9] Jean-Marie Cornuet, Pierre Pudlo, Julien Veyssier, Alexandre Dehne-Garcia, Mathieu Gautier, Raphaël Leblois, Jean-Michel Marin, and Arnaud Estoup. DIYABC v2. 0: a software to make approximate Bayesian computation inferences about population history using single nucleotide polymorphism, DNA sequence and microsatellite data. *Bioinformatics*, 30(8):1187–1189, 2014.

[10] Katalin Csilléry, Michael GB Blum, Oscar E Gaggiotti, and Olivier François. Approximate Bayesian computation (abc) in practice. *Trends in ecology & evolution*, 25(7):410–418, 2010.

[11] Katalin Csilléry, Olivier François, and Michael GB Blum. abc: an R package for approximate Bayesian computation (abc). *Methods in Ecology and Evolution*, 3(3):475–479, 2012.

[12] Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(3):419–474, 2012.

[13] Bruce E Hansen. Nonparametric conditional density estimation. Working paper available at `http://www.ssc.wisc.edu/~bhansen/papers/ncde.pdf`, 2004.

[14] Richard R Hudson. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7(1):44, 1990.

[15] Richard R Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[16] R J Hyndman, D M Bashtannyk, and G K Grunwald. Estimating and visualizing conditional densities. *Journal of Computing and Graphical Statistics*, 5:315–336, December 1996.

[17] Jean-Michel Marin, Louis Raynal, Pierre Pudlo, Mathieu Ribatet, and Christian P Robert. ABC random forests for bayesian parameter inference. *arXiv preprint arXiv:1605.05537*, 2016.

[18] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

[19] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

[20] Jovana Mitrovic, Dino Sejdinovic, and Yee-Whye Teh. DR-ABC: approximate bayesian computation with kernel-based distribution regression. In *International Conference on Machine Learning*, pages 1482–1491, 2016.

[21] Shigeki Nakagome, Kenji Fukumizu, and Shuhei Mano. Kernel approximate Bayesian computation in population genetic inferences. *Statistical applications in genetics and molecular biology*, 12(6):667–678, 2013.

[22] Mijung Park, Wittawat Jitkrittum, and Dino Sejdinovic. K2-ABC: Approximate bayesian computation with kernel embeddings. In *Artificial Intelligence and Statistics*, pages 398–407, 2016.

[23] Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human Y chromosomes: a study of y chromosome microsatellites chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

[24] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[25] Brian D Ripley. Neural networks and related methods for classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 409–456, 1994.

[26] Noah A Rosenberg and Magnus Nordborg. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Reviews Genetics*, 3(5):380–390, 2002.

[27] Emma Saulnier, Olivier Gascuel, and Samuel Alizon. Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. *PLoS computational biology*, 13(3):e1005416, 2017.

[28] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons Sons, 2009.

[29] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[30] Scott A Sisson, Yanan Fan, and Mark M Tanaka. Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.