Approximate Bayesian Computation: A Nonparametric Perspective

Michael G. B. BLUM

Approximate Bayesian Computation is a family of likelihood-free inference techniques that are well suited to models defined in terms of a stochastic generating mechanism. In a nutshell, Approximate Bayesian Computation proceeds by computing summary statistics $\mathbf{s}_{obs}$ from the data and simulating summary statistics for different values of the parameter $\mathbf{\Theta}$. The posterior distribution is then approximated by an estimator of the conditional density $g(\mathbf{\Theta}|\mathbf{s}_{obs})$. In this paper, we derive the asymptotic bias and variance of the standard estimators of the posterior distribution which are based on rejection sampling and linear adjustment. Additionally, we introduce an original estimator of the posterior distribution based on quadratic adjustment and we show that its bias contains a fewer number of terms than the estimator with linear adjustment. Although we find that the estimators with adjustment are not universally superior to the estimator based on rejection sampling, we find that they can achieve better performance when there is a nearly homoscedastic relationship between the summary statistics and the parameter of interest. To make this relationship as homoscedastic as possible, we propose to use transformations of the summary statistics. In different examples borrowed from the population genetics and epidemiological literature, we show the potential of the methods with adjustment and of the transformations of the summary statistics. Supplemental materials containing the details of the proofs are available online.

KEY WORDS: Conditional density estimation; Implicit statistical model; Kernel regression; Local polynomial; Simulation-based inference.

1. INTRODUCTION

Inference in Bayesian statistics relies on the *full posterior distribution* defined as

$$g(\mathbf{\Theta}|D) = \frac{p(D|\mathbf{\Theta})\pi(\mathbf{\Theta})}{p(D)}, \qquad (1)$$

where $\mathbf{\Theta} \in \mathbb{R}^p$ denotes the vector of parameters and D denotes the observed data. The expression given in (1) depends on the *prior distribution* $\pi(\mathbf{\Theta})$, the *likelihood function* $p(D|\mathbf{\Theta})$, and the marginal probability of the data $p(D) = \int_{\mathbf{\Theta}} p(D|\mathbf{\Theta})\pi(\mathbf{\Theta}) \, d\mathbf{\Theta}$. However, when the statistical model is defined in terms of a stochastic generating mechanism, the likelihood can be computationally intractable. Such difficulties typically arise when the generating mechanism involves a high-dimensional variable which is not observed. The likelihood is accordingly expressed as a high-dimensional integral over this missing variable and can be computationally intractable. Methods of inference in the context of these so-called *implicit statistical models* have been proposed by Diggle and Gratton (1984) in a frequentist setting. Implicit statistical models can be thought of as a computer generating mechanism that mimics data generation. In the past 10 years, interests in implicit statistical models have reappeared in population genetics where Beaumont, Zhang, and Balding (2002) gave the name of Approximate Bayesian Computation (ABC) to a family of likelihood-free inference methods.

Since its original developments in population genetics (Fu and Li 1997; Tavaré et al. 1997; Pritchard et al. 1999; Beaumont, Zhang, and Balding 2002), ABC has successfully been applied in a large range of scientific fields such as archaeological science (Wilkinson and Tavaré 2009), ecology (François et al. 2008; Jabot and Chave 2009), epidemiology (Tanaka et al. 2006; Blum and Tran 2010), stereology (Bortot, Coles, and Sisson 2007), or in the context of protein networks (Ratmann et al.

2007). Despite the increasing number of ABC applications, theoretical results concerning its properties are still lacking and the present paper contributes to filling this gap.

In ABC, inference is no more based on the full posterior distribution $g(\mathbf{\Theta}|D)$ but on the *partial* posterior distribution $g(\mathbf{\Theta}|\mathbf{s}_{obs})$ where $\mathbf{s}_{obs}$ denotes a vector of d-dimensional summary statistics computed from the data D. The partial posterior distribution is defined as (Doksum and Lo 1990)

$$g(\mathbf{\Theta}|\mathbf{s}_{obs}) = \frac{p(\mathbf{s}_{obs}|\mathbf{\Theta})\pi(\mathbf{\Theta})}{p(\mathbf{s}_{obs})}. \qquad (2)$$

Of course, the partial and the full posterior distributions are the same if the summary statistics are sufficient with respect to the parameter $\mathbf{\Theta}$.

To generate a sample from the partial posterior distribution $g(\mathbf{\Theta}|\mathbf{s}_{obs})$, ABC with rejection sampling proceeds by simulating n values $\mathbf{\Theta}_i$, $i = 1, \ldots, n$, from the prior distribution π, and then simulating summary statistics $\mathbf{s}_i$ according to $p(\mathbf{s}|\mathbf{\Theta}_i)$. Once the couples $(\mathbf{\Theta}_i, \mathbf{s}_i)$, $i = 1, \ldots, n$, have been obtained, the estimation of the partial posterior distribution is a problem of conditional density estimation. Here we will derive the asymptotic bias and variance of a Nadaraya–Watson type estimator (Nadaraya 1964; Watson 1964), of an estimator with linear adjustment proposed by Beaumont, Zhang, and Balding (2002), and of an original estimator with quadratic adjustment that we propose.

Although replacing the full posterior by the partial one is a crucial approximation in ABC, we will not investigate its consequences here. The reader is referred to Le Cam (1964) and Abril (1994) for theoretical works on the concept of approximate sufficiency; and to Joyce and Marjoram (2008) for a practical method that selects informative summary statistics in ABC. Here, we concentrate on the second type of approximation arising from the discrepancy between the estimated partial posterior distribution and the true partial posterior distribution.

Michael G. B. Blum is Research Associate, CNRS, Laboratoire TIMC-IMAG, Faculté de Médecine, Université Joseph Fourier, Grenoble, France (E-mail: michael.blum@imag.fr).

In this paper, we investigate the asymptotic bias and variance of the estimators of the posterior distribution $g(\theta|\mathbf{s}_{obs})$ ($\theta \in \mathbb{R}$) of a one-dimensional coordinate of $\boldsymbol{\Theta}$. Section 2 introduces parameter inference in ABC. Section 3 presents the main theorem concerning the asymptotic bias and variance of the estimators of the partial posterior. To decrease the bias of the different estimators, we propose, in Section 4, to use transformations of the summary statistics. In Section 5, we show applications of ABC in population genetics and epidemiology.

2. PARAMETER INFERENCE IN ABC

2.1 Smooth Rejection

Assume that the couples $(\boldsymbol{\Theta}_i, \mathbf{s}_i)$, $i = 1, \ldots, n$, have been sampled according to the distribution $p(\mathbf{s}|\boldsymbol{\Theta})\pi(\boldsymbol{\Theta})$. In the context of ABC, the Nadaraya–Watson estimator of the partial posterior mean $E[\boldsymbol{\Theta}|\mathbf{s}_{obs}]$ can be written as

$$m_0 = \frac{\sum_{i=1}^n \boldsymbol{\Theta}_i K_{\mathbf{B}}(\mathbf{s}_i - \mathbf{s}_{obs})}{\sum_{i=1}^n K_{\mathbf{B}}(\mathbf{s}_i - \mathbf{s}_{obs})}, \tag{3}$$

where $K_{\mathbf{B}}(\mathbf{u}) = |\mathbf{B}|^{-1}K(\mathbf{B}^{-1}\mathbf{u})$, $\mathbf{B}$ is the $d \times d$ bandwidth matrix that is assumed to be nonsingular, K is a d-variate kernel such that $\int K(\mathbf{u}) d\mathbf{u} = 1$, and $|\mathbf{B}|$ denotes the determinant of $\mathbf{B}$. Typical choices of kernel encompass spherically symmetric kernels $K(\mathbf{u}) = K_1(\|\mathbf{u}\|)$, in which $\|\mathbf{u}\|$ denotes the Euclidean norm of $\mathbf{u}$ and K_1 denotes a one-dimensional kernel. To estimate the partial posterior distribution $g(\theta|\mathbf{s}_{obs})$ of a one-dimensional coordinate of $\boldsymbol{\Theta}$, we introduce a kernel $\tilde{K}$ that is a symmetric density function on $\mathbb{R}$. Here we will restrict our analysis to univariate density estimation but multivariate density estimation can also be implemented in the same vein. The bandwidth corresponding to $\tilde{K}$ is denoted b' ($b' > 0$) and we use the notation $\tilde{K}_{b'}(\cdot) = \tilde{K}(\cdot/b')/b'$. As the bandwidth b' goes to 0, a simple Taylor expansion shows that

$$E_{\theta'}[\tilde{K}_{b'}(\theta' - \theta)|\mathbf{s}_{obs}] \approx g(\theta|\mathbf{s}_{obs}).$$

The estimation of the partial posterior distribution $g(\theta|\mathbf{s}_{obs})$ can thus be viewed as a problem of nonparametric regression. After substituting $\boldsymbol{\Theta}_i$ by $\tilde{K}_{b'}(\theta_i - \theta)$ in equation (3), we obtain the following estimator of $g(\theta|\mathbf{s}_{obs})$ (Rosenblatt 1969)

$$\hat{g}_0(\theta|\mathbf{s}_{obs}) = \frac{\sum_{i=1}^n \tilde{K}_{b'}(\theta_i - \theta)K_{\mathbf{B}}(\mathbf{s}_i - \mathbf{s}_{obs})}{\sum_{i=1}^n K_{\mathbf{B}}(\mathbf{s}_i - \mathbf{s}_{obs})}. \tag{4}$$

The initial rejection-based ABC estimator consisted of using a kernel K that took 0 or 1 values (Pritchard et al. 1999). This initial method consists simply of rejecting the parameter values for which the simulated summary statistics are too different from the observed ones. Estimation with smooth kernels K was proposed by Beaumont, Zhang, and Balding (2002).

2.2 Regression Adjustment

Besides introducing smoothing in the ABC algorithm, Beaumont, Zhang, and Balding (2002) proposed additionally to adjust the θ_i's to weaken the effect of the discrepancy between $\mathbf{s}_i$ and $\mathbf{s}_{obs}$. In the neighborhood of $\mathbf{s}_{obs}$, they proposed to approximate the conditional expectation of θ given $\mathbf{s}$ by $\hat{m}_1$ where

$$\hat{m}_1(\mathbf{s}) = \hat{\alpha} + (\mathbf{s} - \mathbf{s}_{obs})^t \hat{\boldsymbol{\beta}} \quad \text{for } \mathbf{s} \text{ such that} \quad K_{\mathbf{B}}(\mathbf{s} - \mathbf{s}_{obs}) > 0. \tag{5}$$

The estimates $\hat{\alpha} \in \mathbb{R}$ and $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ are found by minimizing the weighted sum of squared residuals

$$\text{WSSR} = \sum_{i=1}^n \{\theta_i - (\boldsymbol{\alpha} + (\mathbf{s}_i - \mathbf{s}_{obs})^t \boldsymbol{\beta})\}^2 K_{\mathbf{B}}(\mathbf{s}_i - \mathbf{s}_{obs}). \tag{6}$$

The least-squares estimate is given by (Ruppert and Wand 1994)

$$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = (\mathbf{X}^t \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{W} \boldsymbol{\theta}, \tag{7}$$

where $\mathbf{W}$ is a diagonal matrix whose ith element is $K_{\mathbf{B}}(\mathbf{s}_i - \mathbf{s}_{obs})$,

$$\mathbf{X} = \begin{pmatrix} 1 & s_1^1 - s_{obs}^1 & \cdots & s_1^d - s_{obs}^d \\ \vdots & \vdots & \ddots & \vdots \\ 1 & s_n^1 - s_{obs}^1 & \cdots & s_n^d - s_{obs}^d \end{pmatrix}, \qquad \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix},$$

and s_i^j denotes the jth component of $\mathbf{s}_i$. The principle of regression adjustment consists of forming the empirical residuals $\epsilon_i = \theta_i - \hat{m}_1(\mathbf{s}_i)$, and to adjust the θ_i by computing

$$\theta_i^* = \hat{m}_1(\mathbf{s}_{obs}) + \epsilon_i, \qquad i = 1, \ldots, n. \tag{8}$$

Estimation of $g(\theta|\mathbf{s}_{obs})$ is obtained with the estimator of Equation (4) after replacing the θ_i's by the θ_i^*'s. This leads to the estimator proposed by Beaumont, Zhang, and Balding [2002, equation (9)]

$$\hat{g}_1(\theta|\mathbf{s}_{obs}) = \frac{\sum_{i=1}^n \tilde{K}_{b'}(\theta_i^* - \theta)K_{\mathbf{B}}(\mathbf{s}_i - \mathbf{s}_{obs})}{\sum_{i=1}^n K_{\mathbf{B}}(\mathbf{s}_i - \mathbf{s}_{obs})}. \tag{9}$$

To improve the estimation of the conditional mean $m(\mathbf{s})$, we suggest a slight modification to $\hat{g}_1(\theta|\mathbf{s}_{obs})$ using a quadratic rather than a linear adjustment. Adjustment with general nonlinear regression models was already proposed by Blum and François (2010) in ABC. The conditional expectation of θ given $\mathbf{s}$ is now approximated by $\hat{m}_2$ where

$$\hat{m}_2(\mathbf{s}) = \breve{\alpha} + (\mathbf{s} - \mathbf{s}_{obs})^t \breve{\boldsymbol{\beta}} + \tfrac{1}{2}(\mathbf{s} - \mathbf{s}_{obs})^t \breve{\boldsymbol{\gamma}}(\mathbf{s} - \mathbf{s}_{obs})$$

$$\text{for } \mathbf{s} \text{ such that} \quad K_{\mathbf{B}}(\mathbf{s} - \mathbf{s}_{obs}) > 0. \tag{10}$$

The three estimates $(\breve{\alpha}, \breve{\boldsymbol{\beta}}, \breve{\boldsymbol{\gamma}}) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^{d^2}$ are found by minimizing the quadratic extension of the least squares criterion given in (6). Because $\boldsymbol{\gamma}$ is a symmetric matrix, the inference of $\boldsymbol{\gamma}$ only requires the lower triangular part and the diagonal of the matrix to be estimated. The solution to this new minimization problem is given by (7) where the design matrix $\mathbf{X}$ is now equal to

$$\mathbf{X} = \begin{pmatrix} 1 & s_1^1 - s_{obs}^1 & \cdots & s_1^d - s_{obs}^d & \frac{(s_1^1 - s_{obs}^1)^2}{2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & s_n^1 - s_{obs}^1 & \cdots & s_n^d - s_{obs}^d & \frac{(s_n^1 - s_{obs}^1)^2}{2} \\ & (s_1^1 - s_{obs}^1)(s_1^2 - s_{obs}^2) & \cdots & \frac{(s_1^d - s_{obs}^d)^2}{2} \\ & \vdots & \ddots & \vdots \\ & (s_n^1 - s_{obs}^1)(s_n^2 - s_{obs}^2) & \cdots & \frac{(s_n^d - s_{obs}^d)^2}{2} \end{pmatrix}.$$

Letting $\theta_i^{**} = \hat{m}_2(\mathbf{s}_{obs}) + (\theta_i - \hat{m}_2(\mathbf{s}_i))$, the new estimator of the partial posterior distribution is given by

$$\hat{g}_2(\theta|\mathbf{s}_{obs}) = \frac{\sum_{i=1}^n \tilde{K}_{b'}(\theta_i^{**} - \theta)K_{\mathbf{B}}(\mathbf{s}_i - \mathbf{s}_{obs})}{\sum_{i=1}^n K_{\mathbf{B}}(\mathbf{s}_i - \mathbf{s}_{obs})}. \tag{11}$$

Estimators with regression adjustment in the same vein as those proposed in Equations (9) and (11) have already been proposed by Hyndman, Bashtannyk, and Grunwald (1996) and Hansen (2004) for performing conditional density estimation when $d = 1$.

3. ASYMPTOTIC BIAS AND VARIANCE IN ABC

3.1 Main Theorem

To study the asymptotic bias and variance of the three estimators of the partial posterior distribution $\hat{g}_j(\cdot|\mathbf{s}_{obs})$, $j = 0, 1, 2$, we assume that the bandwidth matrix is diagonal $\mathbf{B} = b\mathbf{D}$. A more general result for nonsingular matrix $\mathbf{B}$ is given in the Appendix. In practice, the bandwidth matrix $\mathbf{B}$ may depend on the simulations, but we will assume in this section that it has been fixed independently of the simulations. This assumption facilitates the computations and is classical when investigating the asymptotic bias and variance of nonparametric estimators (Ruppert and Wand 1994).

In the following, the conditional probability distribution of the residual ϵ in the regression model $\theta(\mathbf{s}) = m(\mathbf{s}) + \epsilon$ is denoted $h(\cdot|\mathbf{s})$. The first (resp. second) derivative of a function f with respect to the variable x is denoted f_x (resp. f_{xx}). When the derivative is taken with respect to a vector $\mathbf{x}$, $f_{\mathbf{x}}$ denotes the gradient of f and $f_{\mathbf{xx}}$ denotes the Hessian of f. The variance–covariance matrix of K is assumed to be diagonal and equal to $\mu_2(K)\mathbf{I}_d$. We also introduce the following notations $\mu_2(\tilde{K}) = \int_u u^2 \tilde{K}(u)\, du$, $R(K) = \int_{\mathbf{u}} K^2(\mathbf{u})\, d\mathbf{u}$, and $R(\tilde{K}) = \int_u \tilde{K}^2(u)\, du$. Finally, 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 stays bounded in the limit in probability.

Theorem 1. Assume that $\mathbf{B} = b\mathbf{D}$, in which $b > 0$ is the bandwidth associated to the kernel K, and assume that conditions (A1)–(A5) of the Appendix hold. The bias and variance of the estimators $\hat{g}_j(\cdot|\mathbf{s}_{obs})$, $j = 0, 1, 2$, are given by

$$E[\hat{g}_j(\theta|\mathbf{s}_{obs}) - g(\theta|\mathbf{s}_{obs})] = C_1 b'^2 + C_{2,j} b^2 + O_P((b^2 + b'^2)^2)$$

$$+ O_P\left(\frac{1}{n|\mathbf{B}|}\right), \quad (12)$$

$$\text{Var}[\hat{g}_j(\theta|\mathbf{s}_{obs})] = \frac{C_3}{nb^d b'}(1 + o_P(1)), \quad (13)$$

with

$$C_1 = \frac{\mu_2(\tilde{K}) g_{\theta\theta}(\theta|\mathbf{s}_{obs})}{2},$$

$$C_{2,0} = \mu_2(K)\left(\frac{\mathbf{g_s}(\theta|\mathbf{s})^t_{|\mathbf{s}=\mathbf{s}_{obs}} \mathbf{D}^2 \mathbf{p_s}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})}\right.$$

$$\left. + \frac{\text{tr}(\mathbf{D}^2 \mathbf{g_{ss}}(\theta|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}})}{2}\right), \quad (14)$$

$$C_{2,1} = \mu_2(K)\left(\frac{\mathbf{h_s}(\epsilon|\mathbf{s})^t_{|\mathbf{s}=\mathbf{s}_{obs}} \mathbf{D}^2 \mathbf{p_s}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})}\right.$$

$$+ \frac{\text{tr}(\mathbf{D}^2 \mathbf{h_{ss}}(\epsilon|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}})}{2}$$

$$\left. - \frac{h_\epsilon(\epsilon|\mathbf{s}_{obs})\,\text{tr}(\mathbf{D}^2 \mathbf{m_{ss}}(\mathbf{s}_{obs}))}{2}\right), \quad (15)$$

$$C_{2,2} = \mu_2(K)\left(\frac{\mathbf{h_s}(\epsilon|\mathbf{s})^t_{|\mathbf{s}=\mathbf{s}_{obs}} \mathbf{D}^2 \mathbf{p_s}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})}\right.$$

$$\left. + \frac{\text{tr}(\mathbf{D}^2 \mathbf{h_{ss}}(\epsilon|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}})}{2}\right), \quad (16)$$

and

$$C_3 = \frac{R(K)R(\tilde{K})g(\theta|\mathbf{s}_{obs})}{|\mathbf{D}|p(\mathbf{s}_{obs})}. \quad (17)$$

Remark 1 (Curse of dimensionality). The mean square error (MSE) of an estimator is equal to the sum of its squared bias and its variance. With standard algebra, we find that the MSEs of the three estimators $\hat{g}_j(\cdot|\mathbf{s}_{obs})$, $j = 0, 1, 2$, are minimized when both b and b' are of the order of $n^{-1/(d+5)}$. This implies that the minimal MSEs are of the order of $n^{-4/(d+5)}$. Thus, the rate at which the minimal MSEs converge to 0 decreases as the dimension d of $\mathbf{s}_{obs}$ increases. However, we wish to add words of caution here. First the asymptotic MSE of $n^{-4/(d+5)}$ does not account for the fact that the "constants" C_1, C_2, C_3 involved in Theorem 1 also depend on the dimension of the summary statistics. Second, and more importantly, Scott (1992), in the context of multivariate density estimation, argued that conclusions arising from the same kind of theoretical arguments were in fact much more pessimistic than the empirical evidence. Finally, because the underlying structure of the summary statistics can typically be of dimension lower than d, dimension reduction techniques, such as partial least squares regression or neural networks have been proposed (Wegmann, Leuenberger, and Excoffier 2009; Blum and François 2010).

Remark 2 (Effective local size and effect of design). As shown by Equations (13) and (17), the variance of the estimators can be expressed, up to a constant, as $\frac{1}{\tilde{n}}\frac{g(\theta|\mathbf{s}_{obs})}{b'}$, where the effective local size is $\tilde{n} = n|\mathbf{D}|p(\mathbf{s}_{obs})b^d$. The effective local size is an approximation of the expected number of simulations that fall within the ellipsoid of radii equal to the diagonal elements of $\mathbf{D}$ times b. Thus, Equations (13) and (17) reflect that the variance is penalized by sparser simulations around $\mathbf{s}_{obs}$. Sequential Monte Carlo samplers (Sisson, Fan, and Tanaka 2007; Beaumont et al. 2009; Toni et al. 2009) precisely aim at adapting the sampling distribution of the parameters, a.k.a. the design, to increase the probability of targeting close to $\mathbf{s}_{obs}$. Likelihood-free MCMC samplers have also been proposed to increase the probability of targeting close to $\mathbf{s}_{obs}$ (Marjoram et al. 2003; Sisson and Fan 2010).

Remark 3 (A closer look at the bias). There are two terms in the bias of $\hat{g}_0(\cdot|\mathbf{s}_{obs})$ [Equation (14)] that are related to the smoothing in the space of the summary statistics. The first term in Equation (14) corresponds to the effect of the design and is large when the gradient of $\mathbf{D}p(\cdot)$ is collinear to the gradient of $\mathbf{D}g(\theta|\cdot)$. This term reflects that, in the neighborhood of $\mathbf{s}_{obs}$, there will be an excess of points in the direction of $\mathbf{D}\mathbf{p_s}(\mathbf{s}_{obs})$. Up to a constant, the second term in Equation (14) is proportional to $\text{tr}(\mathbf{D}^2 \mathbf{g_{ss}}(\theta|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}})$ which is simply the sum of the elementwise product of $\mathbf{D}$ and the Hessian $\mathbf{g_{ss}}(\theta|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}}$. This second term shows that the bias is increased when there is more curvature of $g(\cdot|\mathbf{s})$ at $\mathbf{s}_{obs}$ and more smoothing.

For the estimator $\hat{g}_2(\cdot|\mathbf{s}_{obs})$ with quadratic adjustment, the asymptotic bias is the same as the bias of an estimator for which

the conditional mean would be known exactly. Results of the same nature were found, for $d = 1$, by Fan and Yao (1998) when estimating the conditional variance and by Hansen (2004) when estimating the conditional density. Compared to the bias of $\hat{g}_2(\cdot|\mathbf{s}_{obs})$, the bias of the estimator with linear adjustment $\hat{g}_1(\cdot|\mathbf{s}_{obs})$ contains an additional term depending on the curvature of the conditional mean.

3.2 Bias Comparison Between the Estimators With and Without Adjustment

To investigate the differences between the three estimators, we first assume that the partial posterior distribution of θ can be written as $h(\theta - m(\mathbf{s}))$ in which the function h does not depend on $\mathbf{s}$. This amounts to assuming an homoscedastic model in which the conditional distribution of θ given $\mathbf{s}$ depends on $\mathbf{s}$ only through the conditional mean $m(\mathbf{s})$. If the conditional mean m is linear in $\mathbf{s}$, the two constants $C_{2,1}$ and $C_{2,2}$ are null so that the estimators with regression adjustment have a smaller bias than $\hat{g}_0(\cdot|\mathbf{s}_{obs})$. For such ideal models, the bandwidth b of the estimators with regression adjustment can be taken infinitely large so that the variance will be inversely proportional to the total number of simulations n. Still assuming that $g(\theta|\mathbf{s}) = h(\theta - m(\mathbf{s}))$, but with a nonlinear m, the constant $C_{2,2}$ is null so that the estimator $\hat{g}_2(\cdot|\mathbf{s}_{obs})$ has the smallest asymptotic MSE. However, for general partial posterior distributions, it is not possible to rank the three different biases. Consequently, when using the estimators with adjustment, the parameterization of the model should be guided toward making the distributions $g(\theta|\mathbf{s})$ as homoscedastic as possible. To achieve this objective, we propose, in the next section, to use transformations of the summary statistics.

4. CHOOSING A REGRESSION MODEL

4.1 Transformations of the Summary Statistics and the Parameters

To make the regression as homoscedastic as possible, we propose to transform the summary statistics in Equations (5) and (10). Here we consider logarithmic and square root transformations only but a more general family of transformations could also be considered (Box and Tidwell 1962). We choose the transformations that minimize the weighted sum of squared residuals (WSSR) given in Equation (6) in which we take a uniform kernel for the weight function K. The weights $K_{\mathbf{B}}(\mathbf{s}_i - \mathbf{s}_{obs})$ depend on the transformations of the summary statistics and the uniform kernel ensures that the WSSR are comparable for different transformations. Since there are a total of 3^d regression models to consider, greedy algorithm can be considered for large values of 3^d.

Although transformations of the parameter θ in the regression Equations (5) and (10) can also stabilize the variance (Box and Cox 1964), we rather use transformations of θ for guaranteeing that the adjusted parameters θ_i^* and θ_i^{**} lie in the support of the prior distribution (Beaumont, Zhang, and Balding 2002). For positive parameters, we use a log transformation before regression adjustment. After adjusting the logarithm of a positive parameter, we return to the original scale using an exponential transformation. Replacing the logarithm by a logit transformation, we consider the same procedure for the parameters for which the support of the prior is a finite interval.

4.2 Choosing an Estimator of $g(\cdot|\mathbf{s}_{obs})$

In Section 3, we find that there is not a ranking of the three estimators $\hat{g}_j(\cdot|\mathbf{s}_{obs})$, $j = 0, 1, 2$, which is universally valid. Since the three estimators rely on local regressions, of degree 0, 1, and 2, we propose to choose the regression model that minimizes the prediction error of the regression. Because the regression models involve a different number of predictors, we use cross-validation to evaluate the prediction error. We introduce the following leave-one-out estimate

$$CV_j = \sum_{i=1}^{n} (\hat{m}_j^{-i}(\mathbf{s}_i) - \theta_i)^2, \qquad j = 0, 1, 2, \qquad (18)$$

where $\hat{m}_j^{-i}(\mathbf{s}_i)$ denotes the estimate of $m(\theta_i|\mathbf{s}_i)$ obtained, in the neighborhood of $\mathbf{s}_i$, with a local polynomial of degree j by removing the ith point of the training set.

5. EXAMPLES

5.1 Example 1: A Gaussian Model

We are interested here in the estimation of the variance parameter σ^2 in a Gaussian sample. Although Approximate Bayesian Computation is not required for such a simple model, this example will highlight the potential importance of the transformations of the summary statistics and of the methods with adjustment. Assume that we observe a sample of size $N = 50$ in which each individual is a Gaussian random variable $\mathcal{N}(\mu, \sigma^2)$ of mean μ and variance σ^2. We assume a hierarchical prior for μ and σ^2 (Gelman et al. 2003). The prior for σ^2 is an inverse chi-square distribution with one degree of freedom, and the prior for μ is a Gaussian distribution with mean 0 and variance σ^2. We consider the empirical mean $\bar{x}_N$ and variance s_N^2 as the summary statistics. These two statistics are sufficient with respect to the parameter σ^2 (Gelman et al. 2003). The data come from the well-known Iris dataset and consist of the sample of the petal lengths for the virginica species ($\bar{x}_N = 5.552$, $s_N^2 = 0.304$).

We perform a total of 100 ABC replicates. Each replicate consists of simulating $n = 20{,}000$ Gaussian samples. We consider a spherically symmetric kernel for K and an Epanechnikov kernel for K_1. We assume a diagonal bandwidth matrix $\mathbf{B} = b\mathbf{D}$ where $\mathbf{D}$ contains the standard deviation of each summary statistic in the diagonal and b is the 2.5% quantile of the Euclidean distances $\|\mathbf{s}_i - \mathbf{s}_{obs}\|$, $i = 1, \ldots, n$. This procedure amounts to choosing the 500 simulations that provide the best match to the observed summary statistics. In the two following examples, we consider the same number of simulations, the same bandwidth matrix, and the same kernel. Here the true posterior distribution is known exactly (Gelman et al. 2003) and can be compared to the different estimates obtained with ABC. Since σ^2 is a positive parameter, its log is regressed as described in Section 4. As displayed in Figure 1, the estimate with linear adjustment $\hat{g}_1(\sigma^2|\bar{x}_N, s_N^2)$ provides a good estimate provided that the empirical variance is log-transformed in the regression setting. The WSSR criterion selects the right transformation here since it is minimum for the logarithmic transformation in all of the 100 test replicates. When considering $\bar{x}_N$ and $\log s_N^2$ in the regression, both the linear and the quadratic adjustment provide good estimate of σ^2 by contrast to the method without adjustment
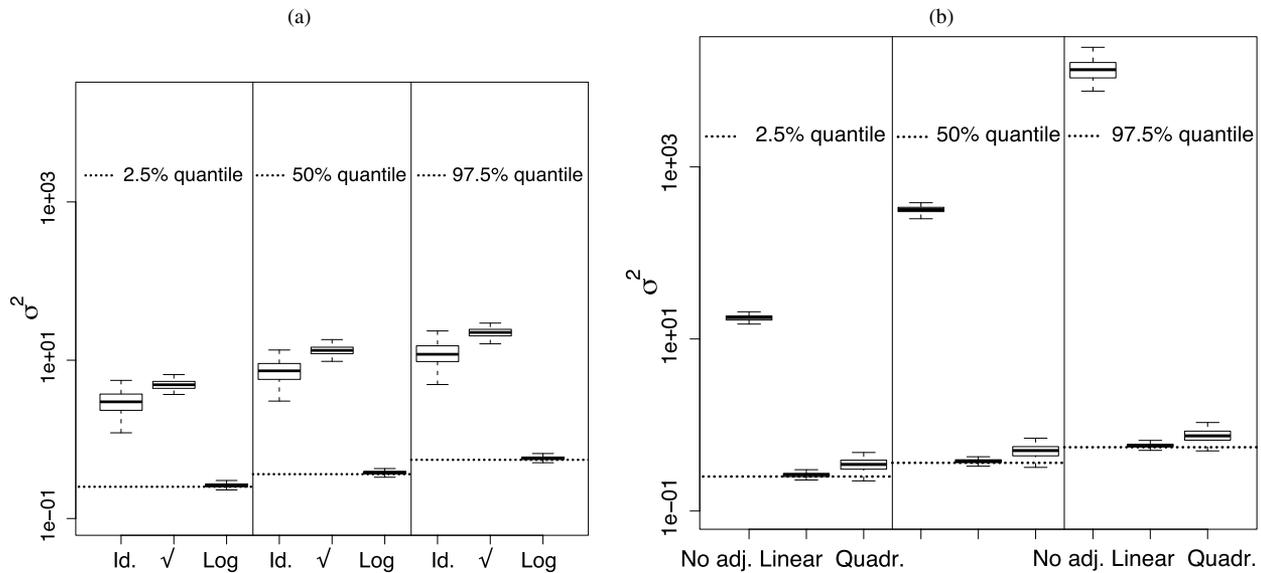
(a)

(b)

Figure 1. Estimation of the posterior quantiles of the variance parameter σ^2 in a Gaussian sample. We perform a total of 100 ABC replicates and we display the boxplots of the estimated posterior quantiles. (a) Estimation of the posterior quantiles with linear adjustment using $(\bar{x}_N, s_N^2)$, $(\bar{x}_N, \sqrt{s_N^2})$, and $(\bar{x}_N, \log s_N^2)$. (b) Estimation of the posterior quantiles with no adjustment and with linear and quadratic adjustment considering $(\bar{x}_N, \log s_N^2)$ as the summary statistics. The horizontal lines correspond to the true posterior quantiles. In this Gaussian example, both log transformation of the empirical variance and regression adjustment are crucial for accurate estimation of the posterior distribution. Id. stands for the identity function, adj. for adjustment, and Quadr. for quadratic.

(see Figure 1). The cross-validation criterion never selects the method without adjustment, selects 74 times linear adjustment and 26 times quadratic adjustment.

5.2 Example 2: Coalescent Model in Population Genetics

ABC was originally developed for inferring parameters of coalescent models in population genetics (Pritchard et al. 1999). Coalescent models describe, in a probabilistic fashion, the tree-like ancestry of genes represented in a sample. Because the ancestral tree is unknown, the likelihood involves an integral over this high-dimensional ancestral tree and is computationally intractable. Here we aim at estimating the Time since the Most Recent Common Ancestor (TMRCA) of a sample of gene. This time is equal to the age of the root of the ancestral tree. A graphical description of the coalescent process and of the TMRCA is given in Figure 2. The coalescent prior for the TMRCA and the whole ancestral tree can be described by the following hierarchical procedure:

1. Simulate the size of the entire population N according to its prior distribution, a uniform distribution between 0 and 10,000 here.
2. Simulate the T_k's, the kth intercoalescence times, as exponential random variables of rate $k(k-1)/(2N)$, $k = 2, \ldots, m$, where m is the number of sequences in the sample. Time is counted in generations here.

The TMRCA is given by the sum of the intercoalescence times $T_2 + \cdots + T_m$. Once the genealogical tree has been generated, DNA sequences are simulated by superimposing mutations along the tree according to a Poisson process of rate u where u is the mutation rate. Here we assume that the mutation

rate is known and we use $u = 1.8 \times 10^{-3}$ mutation/generation for the whole 500 base pairs DNA sequence (Cox 2008). Assuming the *infinitely-many-sites* model, each mutation hit a so-called segregating site that has never been hit before. As summary statistics, we consider the total number of segregating sites S and the mean number of mutations between the ancestor and each individual in the sample. The latter summary statistic is called the ρ statistic and is central in the field of molecular dating (Cox 2008).

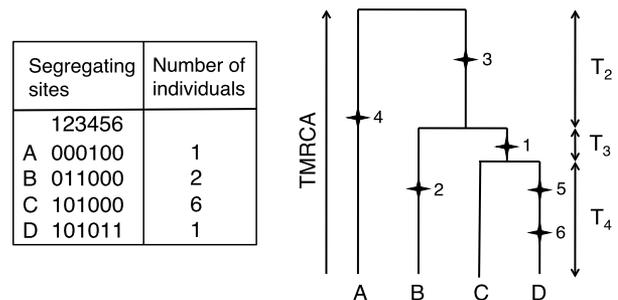| Segregating sites	Number of individuals
123456	
A 000100	1
B 011000	2
C 101000	6
D 101011	1

Figure 2. Coalescent process for simulating DNA sequences. This example is excerpted from Cox (2008). There are a total of 10 DNA sequences. We display only the upper part of the tree in which mutations occur. We omit the lower part corresponding to the coalescence times $T_5, \ldots, T_{10}$. The ancestral sequence is a sequence of 500 base pairs and contains a repetition of 0. The stars denote the $0 \rightarrow 1$ mutations. To generate this tree, a mutation rate of 3.6×10^{-6}/base pairs/generation (equivalent to 1.8×10^{-3}/generation for the 500 bp sequence) was considered. The true TMRCA is equal to 465 generations here. To infer the TMRCA, we consider the number of segregating sites $S = 6$ and the mean number of mutations between the ancestor and the individuals $\rho = 2.10$ as summary statistics.

Table 1. Cross validation criterion for choosing an estimator of the posterior distribution

Parameter	No adjustment	Linear adjustment	Quadratic adjustment
TMRCA (Example 2)	0.90	0.624	**0.620**
Transmission rate $\alpha - \delta$ (Example 3)	1.92	**0.31**	0.34
$R_0 = \alpha/\delta$ (Example 3)	2.15	**1.53**	1.65

We infer the TMRCA using the DNA sequences simulated by Cox (2008). The true TMRCA was equal to 465 generations in his simulation and the values of the summary statistics are $S = 6$ and $\rho = 2.10$. Since the TMRCA is a positive parameter, we use a logarithmic transformation when performing the regression adjustment. The WSSR criterion selects the regression equation $\log \text{TMRCA} = \log \rho + S$ (see Table 1 of the Supplementary Material). The cross validation criterion points to the estimator with quadratic adjustment although the prediction errors obtained with the linear and quadratic regressions are almost the same (see Table 1). In this example, we do not observe the dramatic effect of the transformations and of the adjustments that we found for the Gaussian example. As displayed in Figure 3, both transformations of the summary statistics and regression adjustments do not greatly alter the estimated posterior distribution. Figure 3 also shows that the posterior distribution is clearly more peaked than the prior indicating that the summary statistics convey substantial information about the TMRCA. The 95% credibility interval of the posterior (400–2450) is indeed considerably narrower than the credibility interval of the prior (300–30,800). However, as is typical with molecular dating, there remains considerable uncertainty when

estimating the TMRCA (Cox 2008). The 95% credibility interval of the TMRCA ranges from a value slightly inferior to the true one to a value more than five times larger than the true one.

5.3 Example 3: Birth and Death Process in Epidemiology

To study the rate at which tuberculosis spread in a human population, Tanaka et al. (2006) make use of available genetic data of *Mycobacterium tuberculosis* isolated from different patients. DNA fingerprint at the *IS6110* marker were obtained for 473 isolates sampled in San Francisco during 1991 and 1992 (Small et al. 1994). The *IS6110* fingerprints were grouped into 326 distinct genotypes whose configuration into clusters is represented by

$$30^1 23^1 15^1 10^1 8^1 5^2 4^4 3^{13} 2^{20} 1^{282},$$

where n^k indicates that there are k clusters of size n. To infer the rate of transmission of the disease from this data, Tanaka et al. (2006) introduced a stochastic model of transmission and mutation. We denote by $X_i(t)$ the number of cases of type i at time t, by $G(t)$ the current number of distinct genotypes, and by $N(t)$ the total number of cases. The model starts with $X_1(0) = 1$, $N(0) = 1$, and $G(0) = 1$. We denote by α, δ, and θ, the per-capita birth rate, death rate, and mutation rate. When a birth occurs for an individual of genotype i, the value of $X_i(t)$ is incremented by 1. If the event is a death, the value of $X_i(t)$ is decremented by 1. When a mutation occurs for an individual of genotype i, we assume the *infinitely-many-alleles* model in which a new allele is formed. This means that the value of $X_i(t)$ is decremented by 1 and a case of a new genotype is created. Following Tanaka et al. (2006), the process is stopped when $N = 10,000$. At the stopping time, a sample of size $n = 473$ is drawn from the final population randomly without replacement.
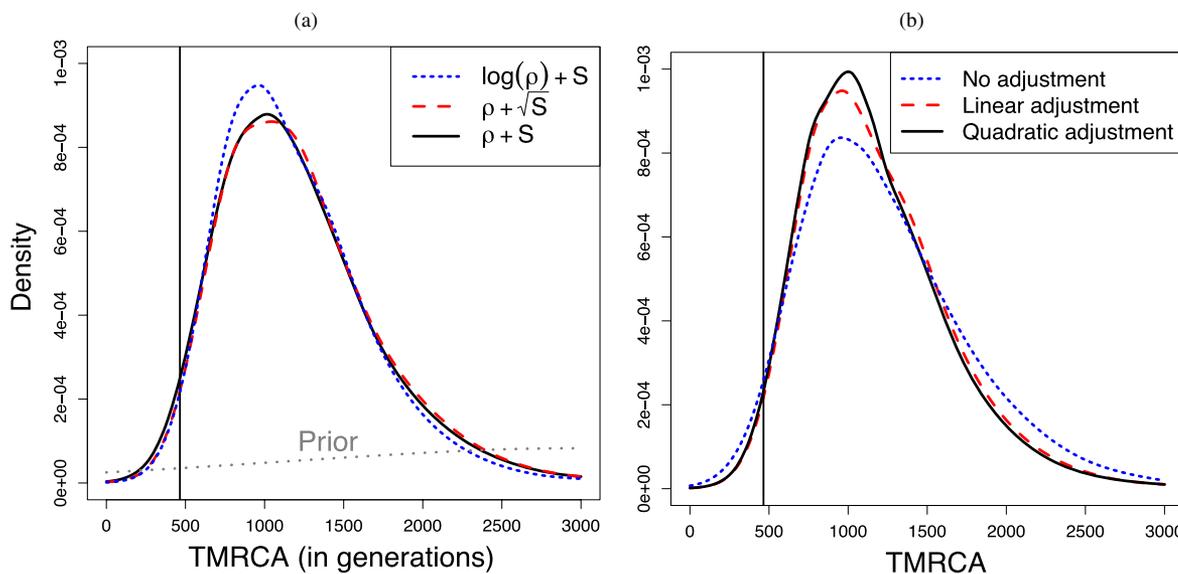

Figure 3. Posterior distribution of the TMRCA. (a) Estimated posterior distributions with linear adjustment considering three different transformations of the summary statistics. The summary statistics $\log \rho$ and S provide the smallest residual error. (b) Estimated posterior distributions using the three different estimates $\hat{g}_j(\text{TMRCA}|(\log \rho, S))$, $j = 0, 1, 2$. The quadratic regression provides the smallest prediction error as found with a leave-one-out estimate. For this coalescent example, both transformations of the summary statistics and regression adjustments do not greatly alter the estimated posterior distribution. The online version of this figure is in color.

As summary statistics, we consider the total number of geno-types G in the sample and the *homozygosity H* of the sample defined as $H = \sum (n_i/n)^2$, where n_i, $i = 1, \ldots, G$, denotes the number of individual of genotype i in the sample. For the San Francisco data, we have $G = 326$ and $H = 1.06\%$. We consider the following prior specification:

$$\theta \sim \mathcal{N}(0.20, 0.07^2), \quad (19)$$

$$\left(\frac{\alpha}{\alpha + \delta + \theta}, \frac{\delta}{\alpha + \delta + \theta}, \frac{\theta}{\alpha + \delta + \theta} \right) \sim \text{Dir}(1, 1, 1),$$

$$\delta < \alpha. \quad (20)$$

The informative prior for θ (in mutation/year) arises from pre-vious estimations of the mutation rate (Tanaka et al. 2006).

We are interested in the estimation of the net transmission rate $\alpha - \delta$, of the doubling time of the disease $\log 2/(\alpha - \delta)$, and of the basic reproduction number $R_0 = \alpha/\delta$. Since they are positive parameters, they are log-transformed in the regression equations. Once log-transformed, the transmission rate and the doubling time are equal up to a multiplicative constant so that the optimal transformation and adjustment are the same for both parameters. We find that transforming G and H with the log function is optimal for inferring the doubling time whereas log-transforming H only is optimal for inferring R_0 (see Table 1 of Supplementary Material). For all parameters, we select linear adjustment based on the cross-validation criterion (see Table 1). As displayed in Figure 4, transformations of the summary sta-tistics and regression adjustments do not greatly alter the es-timated posterior distributions except when estimating R_0. For the transmission rate and the doubling time, the posterior distri-butions greatly differ from the prior distributions (see Figure 4 and Table 2). However, for the reproduction number R_0, the posterior 95% credibility interval is hardly narrower than the prior credibility interval. These comparisons between the prior and the posterior distributions suggest that the genotype data convey much more information for estimating the transmission rate and the doubling time than for estimating the reproduction number R_0. A large credibility interval for the parameter R_0 was also found by Tanaka et al. (2006).

Table 2. Posterior estimates of epidemiological quantities for the San Francisco data

Parameter	Description	95% prior C.I.[a]	Posterior mode	95% posterior C.I.[a]
$\alpha - \delta$	Transmission rate (years)	0.01–9.97	0.56	0.16–0.95
$\log 2/(\alpha - \delta)$	Doubling time (years)	0.06–57.85	1.16	0.73–4.35
α/δ	Reproduction number R_0	1.27–123.32	4.00	2.24–117.45

[a]C.I. stands for credibility intervals.

6. CONCLUSION

In this paper, we presented Approximate Bayesian Computa-tion as a technique of inference that relies on stochastic simula-tions and nonparametric statistics. We introduced an estimator of $g(\theta|\mathbf{s}_{obs})$ based on quadratic adjustment for which the as-ymptotic bias involves fewer terms than the asymptotic bias of the estimator with linear adjustment proposed by Beaumont et al. (2002). More generally, we showed that the bias of the es-timators with regression adjustment [Equations (9) and (11)] is minimal when the distribution of the residual ϵ is independent of $\mathbf{s}$ in the regression model $\theta(\mathbf{s}) = m(\mathbf{s}) + \epsilon$. To make this re-gression model as homoscedastic as possible, we suggested to use transformations of the summary statistics when performing regression adjustment. We proposed to select the transforma-tion of the summary statistics that minimizes the sum of squared residuals within the window of the accepted simulations. In a Gaussian example, we showed that transformations of the sum-mary statistics and regression adjustment can dramatically im-prove inference in ABC. In two other examples borrowed from the population genetics and epidemiology literature, regression adjustment and transformations of the summary statistics had little effect on the estimated posterior distribution. However, above all, these two examples emphasize the potential of ABC for complex models for which the likelihood is not computa-tionally tractable.
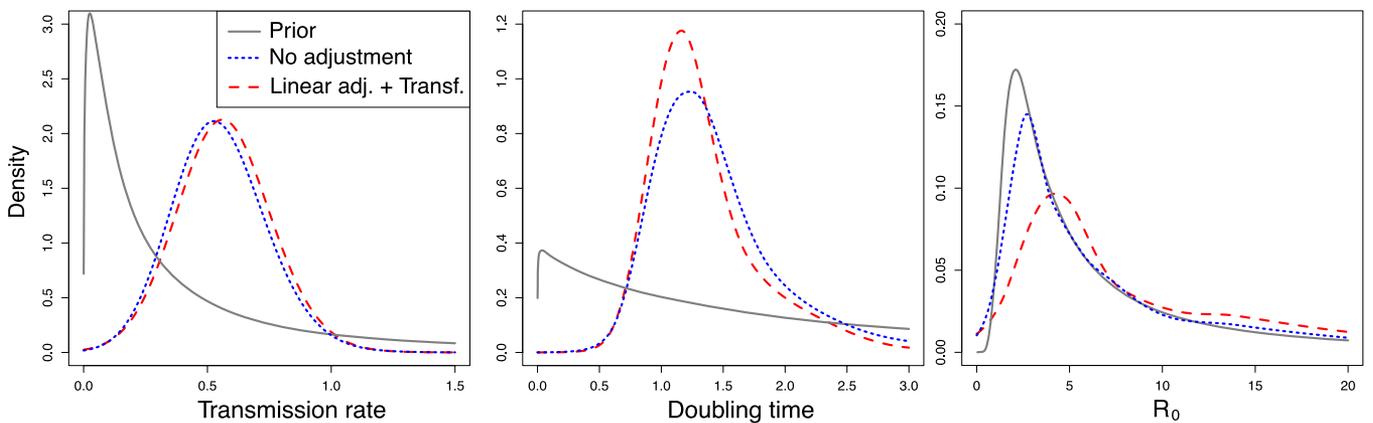

Figure 4. Posterior distributions of key epidemiological quantities for the tuberculosis epidemic in San Francisco. In this example, both transformations of the summary statistics and regression adjustments do not greatly alter the estimated posterior distributions except when estimating R_0. For the transmission rate and the doubling time, the posterior distributions greatly differ from the prior distributions. For the reproduction number R_0, there is not an important difference between the prior and the posterior indicating than the data do not convey enough information for a confident estimation of R_0. The abbreviation transf. stands for transformation. The online version of this figure is in color.

As is expected in nonparametric statistics, we found that the estimators, of the posterior distribution here, suffer from the curse of dimensionality. We found that the rate of convergence of the different estimators is $n^{-4/(d+5)}$ so that it decreases dramatically with the dimension d of the summary statistics. This asymptotic argument gives the impression that just a few summary statistics should be considered in ABC. More generally, it raises the question of the number of summary statistics than can reasonably be handled in ABC. However, there is no simple answer to this difficult question. In a coalescent model, Excoffier, Estoup, and Cornuet (2005) reported good point estimates using as many as $d = 15$ summary statistics. In a related stochastic model in population genetics, Foll, Beaumont, and Gaggiotti (2008) found optimal-point estimates when considering 15 to 25 summary statistics. For ABC practitioners that deal with many summary statistics, a reasonable solution is to consider a technique of dimension reduction (Wegmann, Leuenberger, and Excoffier 2009; Blum and François 2010) and to compare the predictive error [Equation (18)] of the regression estimators obtained with and without dimension reduction.

APPENDIX A: HYPOTHESES OF THEOREM 1

(A1) The kernel K has a finite second-order moment such that $\int \mathbf{u}\mathbf{u}^T K(\mathbf{u})\, d\mathbf{u} = \mu_2(K)\mathbf{I}_d$ where $\mu_2(K) \neq 0$. We also require that all first-order moments of K vanish, that is, $\int \mathbf{u}_i K(\mathbf{u})\, d\mathbf{u} = 0$ for $i = 1, \ldots, d$. As noted by Ruppert and Wand (1994), this condition is fulfilled by spherically symmetric kernels and product kernels based on symmetric univariate kernels.

(A2) The kernel $\tilde{K}$ is a symmetric univariate kernel with finite second-order moment $\mu_2(\tilde{K})$.

(A3) The observed summary statistics $\mathbf{s}_{obs}$ lie in the interior of the support of p. At $\mathbf{s}_{obs}$, all the second-order derivatives of the function p exist and are continuous.

(A4) The point θ is in the support of the partial posterior distribution. At the point $(\theta, \mathbf{s}_{obs})$, all the second-order derivatives of the partial posterior g exist and are continuous. The conditional mean of θ, $m(\mathbf{s})$, exists in a neighborhood of $\mathbf{s}_{obs}$ and is finite. All its second-order derivatives exist and are continuous.

(A5) The sequence of nonsingular bandwidth matrices $\mathbf{B}$ and bandwidths b' is such that $1/(n|\mathbf{B}|b')$, each entry of $\mathbf{B}^t\mathbf{B}$, and b' tend to 0 as $n \to \infty$.

APPENDIX B: PROOF OF THEOREM 1

The three estimators of the partial posterior distribution $\hat{g}_j(\cdot|\mathbf{s}_{obs})$, $j = 0, 1, 2$, are all of the Nadaraya–Watson type. The difficulty in the computation of the bias and variance of the Nadaraya–Watson estimator comes from the fact that it is a ratio of two random variables. Following Pagan and Ullah (1999, p. 98) or Scott (1992), we linearize the estimators in order to compute their biases and variances. We write the estimators of the partial posterior distribution $\hat{g}_j$, $j = 0, 1, 2$, as

$$\hat{g}_j(\theta|\mathbf{s}_{obs}) = \frac{\hat{g}_{j,\mathrm{N}}}{\hat{g}_\mathrm{D}}, \qquad j = 0, 1, 2,$$

where

$$\hat{g}_{0,\mathrm{N}} = \frac{1}{n}\sum_{i=1}^n \tilde{K}_{b'}(\theta_i - \theta)K_\mathbf{B}(\mathbf{s}_i - \mathbf{s}_{obs}),$$

$$\hat{g}_{1,\mathrm{N}} = \frac{1}{n}\sum_{i=1}^n \tilde{K}_{b'}(\theta_i^* - \theta)K_\mathbf{B}(\mathbf{s}_i - \mathbf{s}_{obs}),$$

$$\hat{g}_{2,\mathrm{N}} = \frac{1}{n}\sum_{i=1}^n \tilde{K}_{b'}(\theta_i^{**} - \theta)K_\mathbf{B}(\mathbf{s}_i - \mathbf{s}_{obs}),$$

and

$$\hat{g}_\mathrm{D} = \sum_{i=1}^n K_\mathbf{B}(\mathbf{s}_i - \mathbf{s}_{obs}).$$

To compute the asymptotic expansions of the moments of the three estimators, we use the following lemma:

Lemma 1. For $j = 0, 1, 2$, we have

$$\hat{g}_j(\theta|\mathbf{s}_{obs}) = \frac{E[\hat{g}_{j,\mathrm{N}}]}{E[\hat{g}_\mathrm{D}]} + \frac{\hat{g}_{j,\mathrm{N}} - E[\hat{g}_{j,\mathrm{N}}]}{E[\hat{g}_\mathrm{D}]} - \frac{E[\hat{g}_{j,\mathrm{N}}](\hat{g}_\mathrm{D} - E[\hat{g}_\mathrm{D}])}{E[\hat{g}_\mathrm{D}]^2}$$
$$+ O_P(\mathrm{Cov}(\hat{g}_{j,\mathrm{N}}, \hat{g}_\mathrm{D}) + \mathrm{Var}[\hat{g}_\mathrm{D}]). \quad \text{(B.1)}$$

Proof. Lemma 1 is a simple consequence of a Taylor expansion for the function $(x, y) \to x/y$ in the neighborhood of the point $(E[\hat{g}_{j,\mathrm{N}}], E[\hat{g}_\mathrm{D}])$ (see also Pagan and Ullah 1999). The order of the reminder follows from the weak law of large numbers.

The following lemma gives the asymptotic expansions of all the expressions involved in Equation (B.1):

Lemma 2. Suppose assumptions (A1)–(A5) hold, denote $\epsilon = \theta - m(\mathbf{s}_{obs})$, then we have

$$E[\hat{g}_\mathrm{D}] = p(\mathbf{s}_{obs}) + \frac{1}{2}\mu_2(K)\,\mathrm{tr}(\mathbf{BB}^t\mathbf{p}_\mathbf{ss}(\mathbf{s}_{obs}))$$
$$+ o(\mathrm{tr}(\mathbf{B}^t\mathbf{B})), \quad \text{(B.2)}$$

$$E[\hat{g}_{0,\mathrm{N}}] = p(\mathbf{s}_{obs})g(\theta|\mathbf{s}_{obs}) + \frac{1}{2}b'^2\mu_2(\tilde{K})g_{\theta\theta}(\theta|\mathbf{s}_{obs})p(\mathbf{s}_{obs})$$
$$+ \mu_2(K)\Bigg[\mathbf{g}_\mathbf{s}(\theta|\mathbf{s})^t_{|\mathbf{s}=\mathbf{s}_{obs}}\mathbf{BB}^t\mathbf{p}_\mathbf{s}(\mathbf{s}_{obs})$$
$$+ \frac{1}{2}g(\theta|\mathbf{s}_{obs})\,\mathrm{tr}(\mathbf{BB}^t\mathbf{p}_\mathbf{ss}(\mathbf{s}_{obs}))$$
$$+ \frac{1}{2}p(\mathbf{s}_{obs})\,\mathrm{tr}\big(\mathbf{BB}^t\mathbf{g}_\mathbf{ss}(\theta|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}}\big)\Bigg]$$
$$+ o(b'^2) + o(\mathrm{tr}(\mathbf{B}^t\mathbf{B})), \quad \text{(B.3)}$$

$$E[\hat{g}_{1,\mathrm{N}}] = p(\mathbf{s}_{obs})h(\epsilon|\mathbf{s}_{obs}) + \frac{1}{2}b'^2\mu_2(\tilde{K})h_{\epsilon\epsilon}(\epsilon|\mathbf{s}_{obs})p(\mathbf{s}_{obs})$$
$$+ \mu_2(K)\Bigg[\mathbf{h}_\mathbf{s}(\epsilon|\mathbf{s})^t_{|\mathbf{s}=\mathbf{s}_{obs}}\mathbf{BB}^t\mathbf{p}_\mathbf{s}(\mathbf{s}_{obs})$$
$$+ \frac{1}{2}h(\epsilon|\mathbf{s}_{obs})\,\mathrm{tr}(\mathbf{BB}^t\mathbf{p}_\mathbf{ss}(\mathbf{s}_{obs}))$$
$$+ \frac{1}{2}p(\mathbf{s}_{obs})\,\mathrm{tr}\big(\mathbf{BB}^t\mathbf{h}_\mathbf{ss}(\epsilon|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}}\big)$$
$$- \frac{h_\epsilon(\epsilon|\mathbf{s}_{obs})}{2}\,\mathrm{tr}(\mathbf{BB}^t\mathbf{m}_\mathbf{ss}(\mathbf{s}_{obs}))\Bigg]$$
$$+ o(b'^2) + o(\mathrm{tr}(\mathbf{B}^t\mathbf{B})), \quad \text{(B.4)}$$

$$E[\hat{g}_{2,\mathrm{N}}] = p(\mathbf{s}_{obs})h(\epsilon|\mathbf{s}_{obs}) + \frac{1}{2}b'^2\mu_2(\tilde{K})h_{\epsilon\epsilon}(\epsilon|\mathbf{s}_{obs})p(\mathbf{s}_{obs})$$
$$+ \mu_2(K)\Bigg[\mathbf{h}_\mathbf{s}(\epsilon|\mathbf{s})^t_{|\mathbf{s}=\mathbf{s}_{obs}}\mathbf{BB}^t\mathbf{p}_\mathbf{s}(\mathbf{s}_{obs})$$
$$+ \frac{1}{2}h(\epsilon|\mathbf{s}_{obs})\,\mathrm{tr}(\mathbf{BB}^t\mathbf{p}_\mathbf{ss}(\mathbf{s}_{obs}))$$

$$+ \frac{1}{2} p(\mathbf{s}_{obs}) \, \mathrm{tr}\big(\mathbf{B}\mathbf{B}^t \mathbf{h_{ss}}(\epsilon|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}}\big)\bigg]$$

$$+ o(b'^2) + o(\mathrm{tr}(\mathbf{B}^t\mathbf{B})), \qquad (\text{B.5})$$

$$\mathrm{Var}[\hat{g}_D] = \frac{R(K)p(\mathbf{s}_{obs})}{n|\mathbf{B}|} + O\left(\frac{1}{n}\right) + O\left(\frac{\mathrm{tr}(\mathbf{B}\mathbf{B}^t)}{n|\mathbf{B}|}\right), \qquad (\text{B.6})$$

$$\mathrm{Var}[\hat{g}_{j,N}] = \frac{R(K)R(\tilde{K})g(\theta|\mathbf{s}_{obs})p(\mathbf{s}_{obs})}{nb'|\mathbf{B}|}$$

$$+ O\left(\frac{1}{n}\right) + O\left(\frac{\mathrm{tr}(\mathbf{B}\mathbf{B}^t)}{nb'|\mathbf{B}|}\right) + O\left(\frac{b'}{n|\mathbf{B}|}\right), \qquad (\text{B.7})$$

$$\mathrm{Cov}[\hat{g}_{j,N}, \hat{g}_D] = \frac{R(K)p(\mathbf{s}_{obs})g(\theta|\mathbf{s}_{obs})}{n|\mathbf{B}|}$$

$$+ O\left(\frac{1}{n}\right), \qquad j = 0, 1, 2. \qquad (\text{B.8})$$

Proof. See the Supplemental Material available online.

Theorem 1 is a particular case of the following theorem that gives the bias and variance of the three estimators of the partial posterior distribution for a general nonsingular bandwidth matrix **B**.

Theorem 2. Assume that **B** is a nonsingular bandwidth matrix and assume that conditions (A1)–(A5) hold, then the bias of $\hat{g}_j$, $j = 0, 1, 2$, is given by

$$E[\hat{g}_j(\theta|\mathbf{s}_{obs}) - g(\theta|\mathbf{s}_{obs})] = D_1 b'^2 + D_{2,j} + O_P\big((\mathrm{tr}(\mathbf{B}^t\mathbf{B}) + b'^2)^2\big)$$

$$+ O_P\left(\frac{1}{n|\mathbf{B}|}\right), \qquad j = 0, 1, 2, \quad (\text{B.9})$$

with

$$D_1 = C_1 = \frac{\mu_2(\tilde{K})g_{\theta\theta}(\theta|\mathbf{s}_{obs})}{2},$$

$$D_{2,0} = \mu_2(K)\bigg(\frac{\mathbf{g_s}(\theta|\mathbf{s})^t_{|\mathbf{s}=\mathbf{s}_{obs}}\mathbf{B}\mathbf{B}^t\mathbf{p_s}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})}$$

$$+ \frac{\mathrm{tr}(\mathbf{B}\mathbf{B}^t\mathbf{g_{ss}}(\theta|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}})}{2}\bigg),$$

$$D_{2,1} = \mu_2(K)\bigg(\frac{\mathbf{h_s}(\epsilon|\mathbf{s})^t_{|\mathbf{s}=\mathbf{s}_{obs}}\mathbf{B}\mathbf{B}^t\mathbf{p_s}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})}$$

$$+ \frac{\mathrm{tr}(\mathbf{B}\mathbf{B}^t\mathbf{h_{ss}}(\epsilon|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}})}{2} - \frac{h_\epsilon(\epsilon|\mathbf{s}_{obs})\,\mathrm{tr}(\mathbf{B}\mathbf{B}^t\mathbf{m_{ss}})}{2}\bigg),$$

and

$$D_{2,2} = \mu_2(K)\bigg(\frac{\mathbf{h_s}(\epsilon|\mathbf{s})^t_{|\mathbf{s}=\mathbf{s}_{obs}}\mathbf{B}\mathbf{B}^t\mathbf{p_s}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})} + \frac{\mathrm{tr}(\mathbf{B}\mathbf{B}^t\mathbf{h_{ss}}(\epsilon|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}})}{2}\bigg).$$

The variance of the estimators $\hat{g}_j$, $j = 0, 1, 2$, is given by

$$\mathrm{Var}[\hat{g}_j(\theta|\mathbf{s}_{obs})] = \frac{R(K)R(\tilde{K})g(\theta|\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})n|\mathbf{B}|b'}(1 + o_P(1)). \qquad (\text{B.10})$$

Proof. Theorem 2 is a consequence of Lemmas 1 and 2. Taking expectations on both sides of Equation (B.1), we find that

$$E[\hat{g}_j(\theta|\mathbf{s}_{obs})] = \frac{E[\hat{g}_{j,N}]}{E[\hat{g}_D]} + O_P[\mathrm{Cov}(g_{j,N}, \hat{g}_D) + \mathrm{Var}(\hat{g}_D)]. \quad (\text{B.11})$$

Using a Taylor expansion, and the Equations (B.2)–(B.5), (B.6), and (B.8) given in Lemma 2, we find the bias of the estimators given in Equation (B.9).

For the computation of the variance, we find from Equations (B.1) and (B.11) that

$$\hat{g}_j(\theta|\mathbf{s}_{obs}) - E[\hat{g}_j(\theta|\mathbf{s}_{obs})] = \frac{\hat{g}_{j,N} - E[\hat{g}_{j,N}]}{E[\hat{g}_D]} - \frac{E[\hat{g}_{j,N}](\hat{g}_D - E[\hat{g}_D])}{E[\hat{g}_D]^2}$$

$$+ O_P\left(\frac{1}{n|\mathbf{B}|}\right). \quad (\text{B.12})$$

The order of the reminder follows from Equations (B.6) and (B.8). Taking the expectation of the square of Equation (B.12), we now find

$$\mathrm{Var}[\hat{g}_j(\theta|\mathbf{s}_{obs})] = \frac{\mathrm{Var}[\hat{g}_{j,N}]}{E[\hat{g}_D]^2} + \frac{E[\hat{g}_{j,N}]^2 \,\mathrm{Var}[\hat{g}_D]}{E[\hat{g}_D]^4}$$

$$- 2\,\mathrm{Cov}(\hat{g}_D, \hat{g}_{j,N})\frac{E[\hat{g}_{j,N}]}{E[\hat{g}_D]^3} + o_P\left(\frac{1}{n|\mathbf{B}|b'}\right). \quad (\text{B.13})$$

The variance of the estimators given in Equation (B.10) follows from a Taylor expansion that makes use of Equations (B.2)–(B.8) given in Lemma 2.

SUPPLEMENTAL MATERIALS

Table 1: A table with the weighted sum of squared residuals for each transformation of the summary statistics in Examples 2 and 3. (suppmat_part1.pdf)

Additional proof: Proof of Lemma 2. (suppmat_part2.pdf)

REFERENCES

Abril, J. C. (1994), "On the Concept of Approximate Sufficiency," *Pakistan Journal of Statistics*, 10, 171–177. [1178]

Beaumont, M. A., Marin, J.-M., Cornuet, J.-M., and Robert, C. P. (2009), "Adaptivity for ABC Algorithms: The ABC-PMC Scheme," *Biometrika*, 96, 983–990. [1180]

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002), "Approximate Bayesian Computation in Population Genetics," *Genetics*, 162, 2025–2035. [1178,1179,1181]

Blum, M. G. B., and François, O. (2010), "Non-Linear Regression Models for Approximate Bayesian Computation," *Statistics and Computing*, 20, 63–73. [1179,1180,1185]

Blum, M. G. B., and Tran, V. C. (2010), "HIV With Contact Tracing: A Case Study in Approximate Bayesian Computation," *Biostatistics*, 11, 644–660. [1178]

Bortot, P., Coles, S. G., and Sisson, S. A. (2007), "Inference for Stereological Extremes," *Journal of the American Statistical Association*, 102, 84–92. [1178]

Box, G. E. P., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, Ser. B*, 26, 211–246. [1181]

Box, G. E. P., and Tidwell, P. W. (1962), "Transformation of the Independent Variables," *Technometrics*, 4, 531–550. [1181]

Cox, M. P. (2008), "Accuracy of Molecular Dating With the Rho Statistic: Deviations From Coalescent Expectations Under a Range of Demographic Models," *Human Biology*, 80, 335–357. [1182,1183]

Diggle, P. J., and Gratton, R. J. (1984), "Monte Carlo Methods of Inference for Implicit Statistical Models," *Journal of the Royal Statistical Society, Ser. B*, 46, 193–227. [1178]

Doksum, K. A., and Lo, A. Y. (1990), "Consistent and Robust Bayes Procedures for Location Based on Partial Information," *The Annals of Statistics*, 18, 443–453. [1178]

Excoffier, L., Estoup, A., and Cornuet, J.-M. (2005), "Bayesian Analysis of an Admixture Model With Mutations and Arbitrarily Linked Markers," *Genetics*, 169, 1727–1738. [1185]

Fan, J., and Yao, Q. (1998), "Efficient Estimation of Conditional Variance Functions in Stochastic Regression," *Biometrika*, 85, 645–660. [1181]

Foll, M., Beaumont, M. A., and Gaggiotti, O. (2008), "An Approximate Bayesian Computation Approach to Overcome Biases That Arise When Using Amplified Fragment Length Polymorphism Markers to Study Population Structure," *Genetics*, 179, 927–939. [1185]

François, O., Blum, M. G. B., Jakobsson, M., and Rosenberg, N. A. (2008), "Demographic History of European Populations of *Arabidopsis thaliana*," *PLoS Genetics*, 4 (5), e1000075. [1178]

Fu, Y. X., and Li, W. H. (1997), "Estimating the Age of the Common Ancestor of a Sample of DNA Sequences," *Molecular Biology and Evolution*, 14, 195–199. [1178]

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis* (2nd ed.). *Texts in Statistical Science*, Boca Raton, FL: Chapman & Hall/CRC. [1181]

Hansen, B. E. (2004), "Nonparametric Conditional Density Estimation," working paper, available at *http://www.ssc.wisc.edu/~bhansen/papers/ncde.pdf*. [1180,1181]

Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. (1996), "Estimating and Visualizing Conditional Densities," *Journal of Computing and Graphical Statistics*, 5, 315–336. [1180]

Jabot, F., and Chave, J. (2009), "Inferring the Parameters of the Neutral Theory of Biodiversity Using Phylogenetic Information and Implications for Tropical Forests," *Ecology Letters*, 12, 239–248. [1178]

Joyce, P., and Marjoram, P. (2008), "Approximately Sufficient Statistics and Bayesian Computation," *Statistical Applications in Genetics and Molecular Biology*, 7, Article 26. [1178]

Le Cam, L. (1964), "Sufficiency and Approximate Sufficiency," *The Annals of Mathematical Statistics*, 35, 1419–1455. [1178]

Marjoram, P., Molitor, J., Plagnol, V., and Tavare, S. (2003), "Markov Chain Monte Carlo Without Likelihoods," *Proceedings of the National Academy of Sciences of the United States of America*, 100, 15324–15328. [1180]

Nadaraya, E. (1964), "On Estimating Regression," *Theory of Probability and Applications*, 9, 141–142. [1178]

Pagan, A., and Ullah, A. (1999), *Nonparametric Econometrics*, Cambridge, U.K.: Cambridge University Press. [1185]

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999), "Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites," *Molecular Biology and Evolution*, 16, 1791–1798. [1178,1179,1182]

Ratmann, O., Jørgensen, O., Hinkley, T., Stumpf, M., Richardson, S., and Wiuf, C. (2007), "Using Likelihood-Free Inference to Compare Evolutionary Dynamics of the Protein Networks of *H. pylori* and *P. falciparum*," *PLoS Computational Biology*, 3, e230. [1178]

Rosenblatt, M. (1969), "Conditional Probability Density and Regression Estimates," in *Multivariate Analysis II*, New York: Academic Press, pp. 25–31. [1179]

Ruppert, D., and Wand, M. P. (1994), "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, 22, 1346–1370. [1179,1180, 1185]

Scott, D. W. (1992), *Multivariate Density Estimation*, New York: Wiley. [1180, 1185]

Sisson, S. A., and Fan, Y. (2010), "Likelihood-Free Markov Chain Monte Carlo," in *Handbook of Markov Chain Monte Carlo*, London: Chapman & Hall/CRC Press. [1180]

Sisson, S. A., Fan, Y., and Tanaka, M. (2007), "Sequential Monte Carlo Without Likelihoods," *Proceedings of the National Academy of Sciences of the United States of America*, 104, 1760–1765; errata (2009), 106, 16889. [1180]

Small, P. M., Hopewell, P. C., Singh, S. P., Paz, A., Parsonnet, J., Ruston, D. C., Schecter, G. F., Daley, C. L., and Schoolnik, G. K. (1994), "The Epidemiology of Tuberculosis in San Francisco: A Population-Based Study Using Conventional and Molecular Methods," *New England Journal of Medicine*, 330, 1703–1709. [1183]

Tanaka, M., Francis, A., Luciani, F., and Sisson, S. (2006), "Estimating Tuberculosis Transmission Parameters From Genotype Data Using Approximate Bayesian Computation," *Genetics*, 173, 1511–1520. [1178,1183,1184]

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997), "Inferring Coalescence Times From DNA Sequence Data," *Genetics*, 145, 505–518. [1178]

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., and Stumpf, M. P. (2009), "Approximate Bayesian Computation Scheme for Parameter Inference and Model Selection in Dynamical Systems," *Journal of the Royal Society Interface*, 6, 187–202. [1180]

Watson, G. S. (1964), "Smooth Regression Analysis," *Shankyā, Ser. A*, 26, 359–372. [1178]

Wegmann, D., Leuenberger, C., and Excoffier, L. (2009), "Efficient Approximate Bayesian Computation Coupled With Markov Chain Monte Carlo Without Likelihood," *Genetics*, 182, 1207–1218. [1180,1185]

Wilkinson, R. D., and Tavaré, S. (2009), "Estimating Primate Divergence Times by Using Conditioned Birth-and-Death Processes," *Theoretical Population Biology*, 75, 278–285. [1178]