



Approximate Bayesian Computation: a non-parametric perspective

Michael Blum

► To cite this version:

Michael Blum. Approximate Bayesian Computation: a non-parametric perspective. 2009. hal-00373301v1

HAL Id: hal-00373301

<https://hal.science/hal-00373301v1>

Preprint submitted on 3 Apr 2009 (v1), last revised 8 Apr 2010 (v5)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Approximate Bayesian Computation: a non-parametric perspective

Michael G.B. Blum

CNRS, Laboratoire TIMC-IMAG, Université Joseph Fourier, Grenoble, France

April 3, 2009

Abstract

Approximate Bayesian Computation is a family of likelihood-free inference techniques that are tailored to models defined in terms of a stochastic generating mechanism. In a nutshell, Approximate Bayesian Computation proceeds by computing summary statistics from the data and giving more weight to the values of the parameters for which the simulated summary statistics resemble the observed ones. In this paper, we present Approximate Bayesian Computation as a technique of inference that relies on stochastic simulations and non-parametric statistics. 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 smaller 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. Last, we present model selection in Approximate Bayesian Computation and provide asymptotic properties of two estimators of the model probabilities. As for parameter estimation, the asymptotic results raise the importance of the curse of dimensionality in Approximate Bayesian Computation. Performing numerical simulations, in a simple normal model, confirms that the quality of the estimators deteriorates as the number of summary statistics increases.

Keywords: Conditional density estimation, implicit statistical models, kernel regression, curse of dimensionality, local polynomials

AMS Subject Classification: 62F15, 62G07, 62G08, 62G20, 62M99, 92D10.

Introduction

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

$$g(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{\pi(D)} \quad (1)$$

where θ denotes the vector of interest and D denotes the observed data. The expression given in (1) depends on the *prior distribution* $\pi(\Theta)$, the *likelihood function* $p(D|\Theta)$ and the evidence $p(D) = \int_{\Theta} p(D|\Theta)\pi(\Theta)$. However, for statistical models defined in term of a stochastic generating mechanism, the likelihood can be 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 ten years, interests in implicit statistical models have reappeared in population genetics where Beaumont et al. (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 et al. 2002), ABC has successfully been applied in a large range of scientific fields such as archeological science (Wilkinson and Tavaré 2009), ecology (Jabot and Chave 2009),

epidemiology (Tanaka et al. 2006; Blum and Tran 2009), stereology (Bortot et al. 2007) or for inferring the evolution of protein networks (Ratman et al. 2007). In ABC, inference is no more based on the full posterior distribution $p(\Theta|D)$ but on the *partial* posterior distribution $p(\Theta|\mathbf{s}_{obs})$ where \mathbf{s}_{obs} denotes a vector of d -dimensional summary statistics computed from the data D . The partial posterior distribution is defined as

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

Replacing the likelihood $p(D|\Theta)$ by the partial likelihood $p(\mathbf{s}_{obs}|\Theta)$ is the first approximation inherent to ABC. Of course, the partial and the full posterior distributions are the same if the summary statistics are sufficient with respect to the parameter Θ . Partial posterior distributions have also been considered, in a context that does not involve implicit statistical model, for performing robust Bayesian inference (Doksum and Lo 1990).

Although replacing the full posterior by the partial one is an approximation crucial in ABC, we will not investigate its consequences here. The reader is referred to Le Cam (1964), Abril (1994), Cabrera and Yohai (1999) 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 one $g(\Theta|\mathbf{s}_{obs})$.

In this paper, we investigate the asymptotic bias and variance of the estimator of $g(\Theta|\mathbf{s}_{obs})$ proposed by Beaumont et al. (2002). The originality of the paper is to present ABC as a method of non-parametric statistics for

performing posterior simulation. We additionally propose a new estimator of the partial posterior distribution. The asymptotic bias of this original estimator is shown to contain a smaller number of terms compared to the estimator proposed by Beaumont et al. (2002)

In Section 1, we present the ABC methodology viewed from the angle of non-parametric statistics. In Section 2, we give the main theorems concerning the bias and variance of ABC estimators. Section 3 presents a numerical study in which the properties of the ABC estimators are investigated in a toy model. Section 4 develops the asymptotic theory for model selection in ABC. The proofs of the theorems are given in the Appendix.

1 Approximate Bayesian Computation viewed from the angle of non-parametric statistics

To generate a sample from the partial posterior distribution $g(\Theta|\mathbf{s}_{obs})$, ABC proceeds by simulating n values Θ_i , $i = 1, \dots, n$ from the prior distribution π , and then simulating summary statistics \mathbf{s}_i according to $p(\mathbf{s}|\Theta_i)$. Simulations of the summary statistics according to $p(\mathbf{s}|\Theta_i)$ are feasible because the implicit statistical model is defined in terms of a stochastic generating mechanism. Estimation of the approximate posterior distribution $p(\Theta|\mathbf{s})$ given the couples (Θ_i, \mathbf{s}_i) , $i = 1 \dots n$, is a problem of conditional density estimation and several estimators can be proposed. Note that there exists ABC variants for which the parameters are not sampled from the prior (Marjoram et al. 2003; Sisson et al. 2007; Beaumont et al. 2009; Toni et al. 2009) but we will not investigate their properties here.

1.1 Smooth rejection

In the context of ABC, the estimator of the posterior mean that was originally proposed by Pritchard et al. (1999) is of the Nadaraya-Watson type. The Nadaraya-Watson estimator has been introduced by Nadaraya (1964) and Watson (1964) for inferring the conditional mean of a random variable in a non-parametric fashion (see e.g. Härdle et al. 2004, page 89). In the context of ABC, the Nadaraya-Watson estimator of the partial posterior mean $E[\Theta|\mathbf{s}_{obs}]$ can be written as

$$m_0 = \frac{\sum_{i=1}^n \Theta_i K_B(\mathbf{s}_i - \mathbf{s}_{obs})}{\sum_{i=1}^n K_B(\mathbf{s}_i - \mathbf{s}_{obs})} \quad (2)$$

where B is the *bandwidth matrix* that is assumed to be non-singular, K is a d-variate kernel such that $\int K(\mathbf{u}) d\mathbf{u} = 1$, and $K_B(\mathbf{u}) = |B|^{-1}K(B^{-1}\mathbf{u})$ where $|B|$ denotes the determinant of B .

To go one step further and estimate the approximate posterior distribution $g(\theta|\mathbf{s}_{obs})$ ($\theta \in \mathcal{R}$) of a one-dimensional coordinate of Θ , we introduce a kernel \tilde{K} that is a symmetric density function on \mathcal{R} . Here we will restrict our analysis to univariate density estimation but bivariate 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

$$E_{\theta_0}[K_{b'}(\theta_0 - \theta)|\mathbf{s}_{obs}] \approx g(\theta|\mathbf{s}_{obs}).$$

The estimation of the approximate posterior distribution $g(\theta|\mathbf{s}_{obs})$ can thus be viewed as a problem of nonparametric regression and equation (2) can be

used to estimate $g(\theta|\mathbf{s}_{obs})$ after substituting Θ_i by $\tilde{K}_{b'}(\theta_i - \theta)$. We will refer to this smoother, denoted $\hat{g}_0(\theta)$ as the estimator with smooth rejection for a reason that will be explained below. We have

$$\hat{g}_0(\theta|\mathbf{s}_{obs}) = \frac{\sum_{i=1}^n \tilde{K}_{b'}(\theta_i - \theta) K_B(\mathbf{s}_i - \mathbf{s}_{obs})}{\sum_{i=1}^n K_B(\mathbf{s}_i - \mathbf{s}_{obs})}. \quad (3)$$

In the context of density estimation, this estimator was first proposed by Rosenblatt (1969). In ABC, the choice of \tilde{K} and b' in equation (3) only intervenes when computing the partial posterior distribution at different values of θ as required for graphical display of the posterior. It usually comes in a second step after the sample of the Θ_i 's weighted by the $W_i = K_B(\mathbf{s}_i - \mathbf{s}_{obs})$, $i = 1 \dots n$, has been obtained. For instance, random draws from the partial posterior distribution, as required for Bayesian model checking, are obtained by sampling from the Θ_i 's weighted by the W_i 's (see e.g. Blum and Tran 2009). Note that this is an important difference between standard density estimation in which we are primarily interested in estimating the density at different points whereas ABC aims primarily at simulating replicates from the density $f(\Theta|\mathbf{s}_{obs})$.

1.2 Choice of the bandwidth matrix B

In many applications of ABC (see e.g. Pritchard et al. 1999; Beaumont et al. 2002), the matrix B is a function of a single parameter b , the tolerance threshold, so that $B = bD$ where D is a diagonal matrix in which the j^{th} element measures the scale of the j^{th} summary statistic. In Beaumont et al. (2002), the scales were measured by the standard deviations of each of the components of the summary statistics. This amounts at working with

rescaled summary statistics of variance unity. Interestingly, Cabrera and Yohai (1999), in an unpublished manuscript, developed independently of the population genetics literature an inference method similar to ABC. They chose a bandwidth matrix proportional to $\Sigma^{1/2}$ where Σ is the covariance matrix of the summary statistics. This approach amounts at considering a transformation of the summary statistics so that the transformed summary statistics have an identity covariance matrix.

In Beaumont et al. (2002), the parameter b was chosen so that a prescribed percentage of the simulated summary statistics lie within the support of $K_B(\mathbf{s}_i - \mathbf{s}_{obs})$. Although we just see that the matrix B may depend on the simulations, we will assume in the following theorems 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 non-parametric estimators (Ruppert and Wand 1994).

1.3 Choice of the multivariate kernel K

There are two types of multivariate kernel that are popular for multivariate kernel regression (Härdle et al. 2004, pages 66–70). The first solution is to use a multiplicative kernel

$$K(\mathbf{u}) = K_1(u_1) \dots K_1(u_d), \mathbf{u} = (u_1, \dots, u_d) \in \mathcal{R}^d$$

where K_1 denotes an univariate kernel. Pritchard et al. (1999) considered such a multiplicative kernel with an indicator function $K_1(u) \propto I(|u| < 1)$. The second type of multivariate kernels are the spherically symmetric kernels

and can be obtained by taking

$$K(\mathbf{u}) \propto K_1(\|\mathbf{u}\|),$$

where $\|\cdot\|$ denotes the Euclidean norm.

If K_1 is an indicator function, the spherical kernel amounts at considering a spherical rejection zone above which simulations are rejected and within which all simulations are assigned a weight of 1. Still using an indicator function for K_1 but with a multiplicative kernel for K amounts at defining a rejection zone that is an hypercube rather than an hypersphere. The initial ABC algorithm that consisted of using an indicator function for K_1 (Weiss and Von Haeseler 1998, Pritchard et al. 1999) is known as rejection-based ABC. That is the reason why we refer to the smooth rejection algorithm when considering the smooth estimator given in equation (3). Smooth weighting in the context of ABC was first proposed by Beaumont et al. (2002).

1.4 Regression adjustment

Besides introducing weighting in the ABC algorithm, Beaumont et al. (2002) proposed additionally to adjust the θ_i 's to weaken the effect of the discrepancy between \mathbf{s}_i and \mathbf{s}_{obs} . They proposed to learn the relationship between the expectation of θ denoted as $m(\mathbf{s})$ and \mathbf{s} in the vicinity of \mathbf{s}_{obs} using a polynomial of degree 1. In the neighborhood of \mathbf{s}_{obs} , the conditional expectation of θ given \mathbf{s} is approximated by \hat{m}_1 where

$$\hat{m}_1(\mathbf{s}) = \hat{\alpha} + (\mathbf{s} - \mathbf{s}_{obs})\hat{\beta}^t \text{ for } \mathbf{s} \text{ such that } K_B(\mathbf{s} - \mathbf{s}_{obs}) > 0. \quad (4)$$

The estimates $\hat{\alpha}$ ($\alpha \in \mathcal{R}$) and $\hat{\beta}$ ($\beta \in \mathcal{R}^d$) are found by minimizing the weighted least squares criterion

$$\sum_{i=1}^n \{\theta_i - (\alpha + (\mathbf{s}_i - \mathbf{s}_{obs})^t \beta)\} K_B(\mathbf{s}_i - \mathbf{s}_{obs}). \quad (5)$$

The solution to (5) is given by (Ruppert and Wand 1994, Härdle et al. 2004)

$$(\hat{\alpha}, \hat{\beta}) = (X^t W X)^{-1} X^t W \theta, \quad (6)$$

where W is a diagonal matrix whose i^{th} element is $K_B(\mathbf{s}_i - \mathbf{s}_{obs})$, and

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

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

$$\theta_i^* = \hat{m}_1(\mathbf{s}_{obs}) + \epsilon_i, \quad i = 1, \dots, n. \quad (7)$$

Posterior inference is then based on the θ_i^* 's that have a first moment that match the first moment of $g(\theta|\mathbf{s}_{obs})$ up to the error arising from the estimation of the conditional mean. To generate a sample of the whole-vector Θ according to the partial posterior distribution, the equation (7) can be applied to each coordinate of Θ . This amounts at adjusting the Θ_i by $\Theta_i^* = \hat{M}_1(\mathbf{s}_{obs}) + \xi_i$, where $\xi_i = \Theta_i - \hat{M}_1(\mathbf{s}_i)$, and

$$\hat{M}_1(\mathbf{s}) = (X^t W X)^{-1} X^t W \Theta, \quad \Theta = \begin{pmatrix} \Theta_1^t \\ \vdots \\ \Theta_n^t \end{pmatrix}.$$

Estimation of $g(\theta|\mathbf{s}_{obs})$ is obtained with the Nadaraya-Watson estimator of equation (3) after replacing the θ_i 's by the θ_i^* 's. This leads to the estimator proposed by Beaumont et al. (2002, eq. 9)

$$\hat{g}_1(\theta|\mathbf{s}_{obs}) = \frac{\sum_{i=1}^n \tilde{K}_{b'}(\theta_i^* - \theta) K_B(\mathbf{s}_i - \mathbf{s}_{obs})}{\sum_{i=1}^n K_B(\mathbf{s}_i - \mathbf{s}_{obs})}. \quad (8)$$

Using equation (7), we can express the estimator $\hat{g}_1(\theta|\mathbf{s}_{obs})$ as follows

$$\hat{g}_1(\theta|\mathbf{s}_{obs}) = \frac{\sum_{i=1}^n \tilde{K}_{b'}([\theta_i - \hat{m}_1(\mathbf{s}_i)] - [\theta - \hat{m}_1(\mathbf{s}_{obs})]) K_B(\mathbf{s}_i - \mathbf{s}_{obs})}{\sum_{i=1}^n K_B(\mathbf{s}_i - \mathbf{s}_{obs})}. \quad (9)$$

Noting that $g(\theta|\mathbf{s}_{obs}) = h(\theta - m(\mathbf{s}_{obs})|\mathbf{s}_{obs})$, where $h(\cdot|\mathbf{s}_{obs})$ denotes the conditional density of the residuals, we see that the estimation of $g(\cdot|\mathbf{s}_{obs})$ relies on the estimation of the density h of the residuals. To improve the estimation of the residuals, we suggest a slight modification to the estimator $\hat{g}_1(\theta|\mathbf{s}_{obs})$ using a quadratic rather than a linear adjustment. The conditional expectation of θ given \mathbf{s} is now approximated by \hat{m}_2 where

$$\hat{m}_2(\mathbf{s}) = \tilde{\alpha} + (\mathbf{s} - \mathbf{s}_{obs})^t \tilde{\beta} + \frac{1}{2} (\mathbf{s} - \mathbf{s}_{obs})^t \tilde{\gamma} (\mathbf{s} - \mathbf{s}_{obs}) \text{ for } \mathbf{s} \text{ such that } K_B(\mathbf{s} - \mathbf{s}_{obs}) > 0. \quad (10)$$

The three parameters $(\alpha, \beta, \gamma) \in \mathcal{R} \times \mathcal{R}^d \times \mathcal{R}^{d^2}$ are found by minimizing the quadratic extension of the least square criterion given in (5). Because γ is a symmetric matrix, the inference of γ 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 (6) where the design matrix X is now equal to

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

The new estimator of the partial posterior is given by

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

where

$$\theta_i^{**} = \hat{m}_2(\mathbf{s}_{obs}) + (\theta_i - \hat{m}_2(\mathbf{s}_i)).$$

Note that estimators with regression adjustment such as those proposed in equations (8) and (11) have already been proposed by Hyndman et al. (1996) and Hansen (2004) for performing conditional density estimation when $d = 1$. Hansen (2004) suggested to estimate the conditional mean with a Nadaraya-Watson estimator whereas Hyndman et al. (1996) listed different possible smoothers for estimating the conditional mean. In both papers, they introduced an extra bandwidth parameter for estimating the conditional mean whereas we consider here the same bandwidth B for both the estimation of the conditional mean (equations (4), (5) and (10)) and the estimation of the conditional density (equations (8) and (11)). For $d = 1$, Hyndmann et al. (1996) provided derivations of the asymptotic mean and variance of the Nadaraya-Watson type estimator given by equation (3) and proved its asymptotic normality. Hansen (2004) studied the asymptotic bias, variance and normality of an estimator with regression adjustment for $d = 1$. Note

that other smoothers, related to the Nadaraya-Watson smoother or local polynomials have been proposed for performing conditional density estimation (Fan et al. 1996; De Gooijer and Zerom 2002).

2 Bias and variance of the estimators of the conditional mean

In this section, we study the asymptotic bias and variance of the estimators of the partial posterior distribution $\hat{g}_j(\cdot|\mathbf{s}_{obs})$, $j = 0, 1, 2$, given by equations (3), (8), and (11). We assume, in the following, that the bandwidth matrix is diagonal $B = bD$. A more general result for non-singular matrix B is given in the Appendix. The first (resp. second) derivative of a function f with respect 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 $\mu_2(K)I_d$. We additionally 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 $R(\tilde{K}) = \int_u \tilde{K}^2(u) du$.

Theorem 1 *Assume that $B = bD$ and assume that conditions (A1):(A5) of the Appendix hold. The bias of the estimators $\hat{g}_j(\cdot|\mathbf{s}_{obs})$, $j = 0, 1, 2$, is given by*

$$E[\hat{g}_j(\theta|\mathbf{s}_{obs}) - g(\theta|\mathbf{s}_{obs})] = C_1 b'^2 + C_{2,j} b^2 + O_P((b+b')^2) + O_P\left(\frac{1}{n|B|}\right), \quad j = 0, 1, 2, \quad (12)$$

with

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

$$C_{2,0} = \mu_2(K) \left(\frac{g_{\mathbf{s}}(\theta|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}}^t D^2 p_{\mathbf{s}}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})} + \frac{\text{tr}(D^2 g_{\mathbf{ss}}(\theta|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}})}{2} \right), \quad (13)$$

$$C_{2,1} = \mu_2(K) \left(\frac{h_{\mathbf{s}}(\epsilon|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}}^t D^2 p_{\mathbf{s}}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})} + \frac{\text{tr}(D^2 h_{\mathbf{ss}}(\epsilon|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}})}{2} - \frac{h_{\epsilon}(\epsilon|\mathbf{s}_{obs}) \text{tr}(D^2 m_{\mathbf{ss}})}{2} \right), \quad (14)$$

and

$$C_{2,2} = \mu_2(K) \left(\frac{h_{\mathbf{s}}(\epsilon|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}}^t D^2 p_{\mathbf{s}}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})} + \frac{\text{tr}(D^2 h_{\mathbf{ss}}(\epsilon|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}})}{2} \right). \quad (15)$$

The variance of the estimators $\hat{g}_j(\cdot|\mathbf{s}_{obs})$, $j = 0, 1, 2$, is given by

$$\text{Var}[\hat{g}_j(\theta|\mathbf{s}_{obs})] = \frac{C_3}{nb^d b'} (1 + o_P(1)), \quad j = 0, 1, 2, \quad (16)$$

where

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

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 are minimized when both b and b' are of the order of $n^{-1/(d+5)}$. This implies that the MSEs are of the order of $n^{-4/(d+5)}$. Thus, the rate at which the MSEs converge to 0 decreases importantly as the dimension of \mathbf{S} increases. This phenomenon known as the *curse of dimensionality* is a particular acute issue for the three estimators given by equations (3), (8), and (11).

Remark 2. Effect of design Both the constants $C_{2,0}$, $C_{2,1}$, and $C_{2,2}$ involved in the biases of the estimators, and the variance of the estimators are inversely proportional to the partial evidence $p(\mathbf{s}_{obs})$. In the terminology of

non-parametric statistics, the estimators of the partial posterior distribution are not design-adaptative (Fu 1992). This implies that estimating $g(\theta|\mathbf{s}_{obs})$ will more cumbersome when the probability of observing \mathbf{s}_{obs} is small. The choice of the prior is, therefore, of considerable importance here since the MSEs will be large if the prior is chosen so that partial evidence $p(\mathbf{s}_{obs})$ is small. Note that the sequential Monte Carlo samplers (Sisson et al. 2007, Beaumont et al. 2009, Toni et al. 2009) precisely aims at adapting the sampling distribution of the parameters, i.e. the design, to increase the probability of targeting close to \mathbf{s}_{obs} .

Remark 3. A closer look at the bias There are two terms in the bias of $\hat{g}_0(\cdot|\mathbf{s}_{obs})$ that are related to the smoothing in the space of the summary statistics. The first term in equation (13) corresponds to the effect of the design and is large when the gradient of $Dg(\theta|\cdot)$ is collinear to the gradient of $Dp(\cdot)$. This term reflects that, in the neighborhood of \mathbf{s}_{obs} , there will be an excess of points in the direction of $Dp_{\mathbf{s}}(\mathbf{s}_{obs})$. Concerning the second term in equation (13), we note that $\text{tr}(D^2 g_{ss}(\theta|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}})$ is simply the sum of the elementwise product of D and the hessian $g_{ss}(\theta|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}}$. For minimizing this term, it is thus optimal to have small values of D_j for the directions j for which the curvature of $g(\theta|\cdot)$ is large.

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 by Fan and Yao (1998) when estimating the conditional variance and Hansen (2004) when estimating the conditional density for $d = 1$. For the estimator with linear

adjustment $\hat{g}_1(\cdot|\mathbf{s}_{obs})$, there is an additional term related to the curvature of the conditional mean.

Remark 4. Effective local size As shown by equations (16) and (17), the variance of the estimators can be expressed, up to a constant, as $\frac{1}{\tilde{n}} \frac{g(\theta|\mathbf{s}_{obs})}{b^d}$, where the effective local size is $\tilde{n} = n|D|p(\mathbf{s}_{obs})b^d$. 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 D times b . Thus equations (16) and (17) reflects that the variance is penalized by larger partial posterior values and sparser simulations around \mathbf{s}_{obs} (Ruppert and Wand 1994). As usual for smoothing problem, there is a bias-variance tradeoff since large bandwidths b give large bias and small variance whereas small bandwidths b give large variance and small bias.

Remark 4. Bias comparison 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 at assuming an homoscedastic model in which the conditional distribution of θ given s depends on \mathbf{s} only through the conditional mean $m(\mathbf{s})$. If the conditional mean m is linear in \mathbf{s} , both $C_{2,1}$ and $C_{2,2}$ are null involving that the estimators with regression adjustment have a smaller bias than the smooth rejection estimator. For such ideal models, the bandwidth b of the estimators with regression adjustment can be taken sufficiently large so that the variance of the estimators will be small. Still assuming that $g(\theta|\mathbf{s}) = h(\theta - m(\mathbf{s}))$, but with a non-linear 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 and the choice of the summary statistics shall be guided toward making the model $\theta = m(\mathbf{s})$ as homoscedastic as possible. This explains why the Box-Cox transformations (Box and Cox 1964) are usually considered in ABC before making regression adjustment. Comparison of asymptotic biases for regression and non-regression based conditional density estimators can also be found in Hansen (2004) for $d = 1$.

3 Numerical comparison between the estimators

In this section, we consider a simple normal model to illustrate the curse of dimensionality in ABC. We assume that the d -dimensional data are drawn according to a normal distribution with mean $\mu = (\mu_1, \dots, \mu_D)$ and a variance-covariance matrix equal to the $d \times d$ identity matrix I_d . The prior for μ is a d -dimensional gaussian distribution of mean μ_0 and variance-covariance matrix I_d . Here, we study the properties of the estimators of the distribution of e^{μ_1} based on the summary statistics consisting of the d -dimensional empirical mean $\bar{\mathbf{x}}$ of the data. Note that the empirical mean of the first component of the data \bar{x}_1 is a sufficient statistic with respect to e^{μ_1} so that the partial posterior is the same as the full posterior in this example. The $d - 1$ additional empirical means convey no information for estimating e^{μ_1} and are added here to show that incorporating useless summary statistics can have a dramatic effect on the estimation of the posterior. We assume in the following that

$\bar{\mathbf{x}}_i = 0$, for $i = 1, \dots, d$ and that the sample size is $M = 10$. In this simple toy model, the posterior distribution is known and is a log-normal distribution with mean and variance (on a log scale) $-\mu_0/(M+1)$, and $1/(M+1)$. We choose a spherically symmetric kernel for K so that $K(\mathbf{u}) = K_1(\|\mathbf{u}\|)$ and we consider the kernel K_1 for \tilde{K} where K_1 denotes the Epanechnikov kernel.

3.1 How many simulations are required to reach a given level of accuracy

Here we compute the minimum number n_{\min} of simulations that are required to reach a given level of accuracy when estimating the parameter e^{μ_1} . The number n_{\min} is defined as the smallest number of simulations so that the relative squared error is less than 10% when estimating the posterior distribution at 0. Similar computations were performed by Silverman (1986) to illustrate the curse of dimensionality for density estimation. The mean square errors were computed using equations (12) and (16) in which the constants can be analytically derived in this simple example. The optimal bandwidths for b and b' were found by numerical minimization of the asymptotic mean square errors. To simplify the computations, we assume that the prior for μ is a gaussian distribution of mean $\mu_0 = 0$. As displayed by Figure 1, the regression-based estimators require a significantly smaller number of simulations for $d \leq 6$ but the improvement becomes negligible for $d > 6$. To give a quantitative flavor of the importance of the curse of dimensionality, we note that a minimum number of approximately one million of simulations is required to have a relative mean square error smaller than 10% when

$d = 6$ and this number increases to more than a thousand of billions when $d = 10$. Compared to the estimator with linear adjustment, the estimator with quadratic adjustment requires less simulation when $d = 1, 3, 4$ but this is not true for $d = 2$. When the dimension is larger than 5, there is no more significant differences between the two estimators with adjustment.

Note that estimating e^{μ_1} rather μ_1 is a really loose parameterization here because the regression-based estimators would manage to cope with the curse of dimensionality when estimating μ_1 . Indeed the model $\mu_1 = m(\bar{\mathbf{x}}) + \epsilon$ is linear and homoscedastic so that the term involving b^2 in the bias (equation (12)) would be null for the regression-based estimators. The (loose) parameterization that has been chosen here illustrates 1) the importance of parameterization in ABC and 2) that the regression-based estimators can typically cope with the curse of dimensionality for intermediate values of d but will be inefficient for large values of d .

3.2 Comparison between the estimators based on simulations

To further investigate the differences between the three estimators, we compare the three ABC estimators \hat{g}_j , $j = 0, 1, 2$, to the true posterior distribution of e^{μ_1} . Here we set $\mu_0 = 1$. To compare the different estimators, we compute the mean integrated square error (MISE) defined as

$$\text{MISE} = E \left[\int_{\theta} \{g(\theta|\mathbf{s}_{obs}) - \hat{g}_j(\theta|\mathbf{s}_{obs})\}^2 d\theta \right], \quad j = 0, 1, 2,$$

in which averaging is done with respect to the vector of simulations (θ_i, \mathbf{s}_i) , $i = 1, \dots, n$. The integrated square errors are averaged using a total of 500

replicates, each consisting of performing $n = 10,000$ simulations. We choose a diagonal bandwidth matrix $B = bD$ where the diagonal elements of D contain the mean absolute deviations of each component of the summary statistics. For choosing the bandwidth b , we follow a common practice in the ABC literature consisting of choosing *a priori* the percentage of accepted simulations, i.e. the percentage of simulations for which $K_B(\mathbf{s}_i - \mathbf{s}_{obs})$ is different from 0. Here the percentage of accepted simulations is set to 5%. Note that many alternative methods have been considered for the choice of the bandwidth b when performing conditional density estimation or estimation of the conditional distribution function (Fan and Yim 2004, Hall et al. 1999, Hall et al. 2004). The bandwidth b' is computed using the Silverman's rule of thumb (Silverman 1986, page 48, eq. (3.31)). The integration required to compute the MISE is performed using a simple trapezoidal integration rule with 512 points equally spaced points between 0 and 3.

As displayed by Figure 2 and in accordance with Theorem 1, the MISE increases as the dimension of the summary statistics increases. For $d = 1$, the MISEs of the three estimators are the same and the improvement achieved by the estimators with regression-adjustment appears when $d \geq 2$. When $3 \leq d \leq 8$, the estimators with quadratic regression adjustment achieves the lowest MISE but the situation reverses for $d \geq 9$ where the lowest MISE is achieved by the estimator with linear adjustment. This can be explained by the variance of the estimator with quadratic adjustment that can be important for large values of d . Indeed, the estimator with quadratic adjustment requires the inference of $d(d+3)/2 + 1$ parameters

whereas the estimator with linear adjustment requires the inference of only $d + 1$ parameters. Note that the potential differences between the variances of the different estimators is not captured by the first-order derivation of the asymptotic variances obtained in Theorem 1.

4 Model selection

4.1 Estimation of the probability of models

For sake of simplicity, we assume that there are two competitive models \mathcal{M}_1 and \mathcal{M}_2 that are *a priori* equally likely. If we denote by $p_1(\mathbf{s}_{obs})$ and $p_2(\mathbf{s}_{obs})$ the partial evidence in each model, the probability of the first model is defined as

$$p(\mathcal{M}_1|\mathbf{s}_{obs}) = \frac{p_1(\mathbf{s}_{obs})}{p_1(\mathbf{s}_{obs}) + p_2(\mathbf{s}_{obs})}.$$

Alternatively, Bayesian model selection may be based on the the partial Bayes factor BF_{12} of \mathcal{M}_1 versus \mathcal{M}_2 defined as the odds in favor of \mathcal{M}_1

$$\text{BF}_{12} = \frac{p_1(\mathbf{s}_{obs})}{p_2(\mathbf{s}_{obs})}.$$

Inference of the model probabilities and partial Bayes factors are straightforward in ABC and have been considered for discriminating different models of human evolution for instance (Fagundes et al. 2007). In the same vein as in Section 3, we do not study the error arising from the difference between the partial Bayes factor and the *full* one $p_1(D)/p_2(D)$ but we focus on the error arising from the estimation of the partial Bayes factor. We assume here that either $n/2$ simulations have been performed in each model or that the generative models \mathcal{M}_1 and \mathcal{M}_2 were chosen with equal probability for

each simulation. We denote by Y_i , $i = 1, \dots, n$ an indicator variable equal to 1 if the i^{th} simulation was performed using the generative model of \mathcal{M}_1 and 0 otherwise. An estimator of $p(\mathcal{M}_1|\mathbf{s}_{obs})$ is obtained using the following Nadaraya-Watson estimator

$$\hat{p}_0(\mathcal{M}_1|\mathbf{s}_{obs}) = \frac{\sum_{i=1}^n Y_i K_B(\mathbf{s}_i - \mathbf{s}_{obs})}{\sum_{i=1}^n K_B(\mathbf{s}_i - \mathbf{s}_{obs})}. \quad (18)$$

Using equation (18) to compute the partial Bayes factor, we get

$$\hat{\text{BF}}_0 = \frac{\sum_{i=1}^n Y_i K_B(\mathbf{s}_i - \mathbf{s}_{obs})}{\sum_{i=1}^n (1 - Y_i) K_B(\mathbf{s}_i - \mathbf{s}_{obs})}. \quad (19)$$

When an indicator kernel is considered for K , the partial Bayes factor is simply estimated as the ratio of the acceptance rates in each model. Pritchard et al. (1999) considered the ratio of the acceptance rates as an estimate of the Bayes factor whereas François et al. (2008) and Wilkinson (2009) proposed to estimate the Bayes factor with the more generic ratio given by equation (19).

An alternative method has been proposed by Beaumont (2008) to estimate the model probabilities in ABC. Viewing the estimation problem in a regression setting in which \mathbf{s} is the predictive variable and Y is the indicator variable to predict, Beaumont (2008) proposed to use local logistic regression to estimate $E[Y|\mathbf{s}_{obs}] = p(\mathcal{M}_1|\mathbf{s}_{obs})$. In local logistic regression, the log-odds of the model probabilities are approximated by a linear function so that $\log[p(\mathcal{M}_1|\mathbf{s})/(1 - p(\mathcal{M}_1|\mathbf{s}))] = \delta_0 + (\mathbf{s} - \mathbf{s}_{obs})^t \delta_1$. The log of the weighted likelihood \mathcal{L} can be written as

$$\mathcal{L}(\delta_0, \delta_1) = \sum_{i=1}^n \{Y_i \log(g_{\delta_0, \delta_1}(\mathbf{s}_i)) + (1 - Y_i) \log(1 - g_{\delta_0, \delta_1}(\mathbf{s}_i))\} K_B(\mathbf{s}_i - \mathbf{s}_{obs}), \quad (20)$$

where $g_{\delta_0, \delta_1}(\mathbf{s}) = e^{\delta_0 + (\mathbf{s} - \mathbf{s}_{obs})^t \delta_1} / (1 + e^{\delta_0 + (\mathbf{s} - \mathbf{s}_{obs})^t \delta_1})$. Denoting by $\hat{\delta}_0$ and $\hat{\delta}_1$, the parameters that maximize the log-likelihood given in equation (20), the probability of model \mathcal{M}_1 is estimated as

$$\hat{p}_1(\mathcal{M}_1 | \mathbf{s}_{obs}) = \frac{e^{\hat{\delta}_0}}{1 + e^{\hat{\delta}_0}}. \quad (21)$$

The optimization of equation (20) has no explicit solution and iterative algorithms such as iteratively reweighted least squares shall be considered (McCullagh and Nelder 1989).

In the following, we give the main theorem concerning the bias and variance of both estimators $\hat{p}_j(\mathcal{M}_1 | \mathbf{s}_{obs})$, $j = 0, 1$. We assume here that the bandwidth matrix is diagonal $B = bD$ but a more general theorem for non singular bandwidth matrix B could also be obtained from Fan et al. (1995).

Theorem 2 *Assume that conditions (A1) and (A5') of the appendix hold and that the condition (A3) holds for both \mathcal{M}_1 and \mathcal{M}_2 . The bias of the estimators $\hat{p}_0(\mathcal{M}_1 | \mathbf{s}_{obs})$ and $\hat{p}_1(\mathcal{M}_1 | \mathbf{s}_{obs})$ is given by*

$$E[\hat{p}_j(\mathcal{M}_1 | \mathbf{s}_{obs}) - p(\mathcal{M}_1 | \mathbf{s}_{obs})] = \left(\mu_2(K) \frac{\text{tr}(D^2 p_{\mathbf{ss}}(\mathcal{M}_1 | \mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}})}{2} + E_j \right) b^2, \quad (22)$$

for $j = 0, 1$, where

$$E_0 = \mu_2(K) \frac{p_{\mathbf{s}}(\mathcal{M}_1 | \mathbf{s})^t|_{\mathbf{s}=\mathbf{s}_{obs}} D^2 p_{\mathbf{s}}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})},$$

$$E_1 = \frac{\mu_2(K)}{2} \left(p_{\mathbf{s}}(\mathcal{M}_1 | \mathbf{s})^t|_{\mathbf{s}=\mathbf{s}_{obs}} D^2 p_{\mathbf{s}}(\mathcal{M}_1 | \mathbf{s}) \left(\frac{1}{1 - p(\mathcal{M}_1 | \mathbf{s}_{obs})} - \frac{1}{p(\mathcal{M}_1 | \mathbf{s}_{obs})} \right) \right),$$

and $p(\mathbf{s}) = \frac{p_1(\mathbf{s})}{2} + \frac{p_2(\mathbf{s})}{2}$. The variance of the estimators is given by

$$\text{Var}[\hat{p}_j] = \frac{1}{nb^d} \frac{\mu_2(K)p(\mathcal{M}_1|\mathbf{s}_{obs})(1 - p(\mathcal{M}_1|\mathbf{s}_{obs}))}{|D|p(\mathbf{s}_{obs})}, \quad j = 0, 1.$$

Proof. The estimator given in equation (18) is a Nadaraya-Watson estimator and its asymptotics are given by the standard asymptotic bias and variance of the Nadaraya-Watson estimator (Härdle et al. 2004, page 131). The asymptotic bias and variance of the estimator $\hat{p}_1(\mathcal{M}_1|\mathbf{s}_{obs})$ given in equation (21) can be obtained from the Theorem 3 of Fan et al. (1995) dedicated to multivariate local regression for generalized linear models. \square

Remark 1 Curse of dimensionality Standard algebra shows that the optimal bandwidths are found when $b \propto n^{-1/(d+4)}$ for which the mean square error is of the order of $n^{-4/(d+4)}$.

Remark 2 Effect of design As for parameter estimation, it is not possible to give an universal ranking of the two different estimators. However, we find that the estimator based on local logistic regression is design-adaptive. By contrast, the bias of the Nadaraya-Watson estimator is large when the averaged partial evidence $p(\mathbf{s}) = \frac{p_1(\mathbf{s}_{obs})}{2} + \frac{p_2(\mathbf{s}_{obs})}{2}$ is small.

Remark 3 Bayes factor The estimators of the logarithm of the Bayes factor are simply obtained as $\hat{\text{BF}}_j = \phi(\hat{p}_j)$, for $j = 0, 1$, where $\phi(x) = \log(x/(1-x))$. Using a Taylor expansion, we find that the bias and variance of $\hat{\text{BF}}_j$ have the same asymptotic behavior as those obtained in Theorem 2 except that the bias is multiplied by $\phi'(\hat{p}_j) = 1/((\hat{p}_j)(1-\hat{p}_j))$ and that the variance is multiplied by $\phi'(\hat{p}_j)^2$. When estimating the log of the Bayes factor using local logistic regression, the bias takes a simple form as (see also Fan

et al. 1995, Theorem 3)

$$E[\hat{\text{BF}}_1 - \text{BF}] = \frac{\mu_2(K)}{2} \text{tr}(D^2 \psi_{\mathbf{ss}}(\mathbf{s}_{obs})) b^2,$$

where $\psi(\mathbf{s}) = \phi(\hat{p}_1(\mathbf{s}))$.

4.2 Numerical example

We consider both the Nadaraya-Watson estimator $\hat{p}_0(\mathcal{M}_1|\mathbf{s}_{obs})$ and the estimator with local logistic regression $\hat{p}_1(\mathcal{M}_1|\mathbf{s}_{obs})$ in the context of the normal model given in Section 3. The maximization of the weighted likelihood (equation (20)) was performed using the R routine `glm` (R Core Team 2008). Here we posit that model \mathcal{M}_1 assumes that μ_1 is equal to 0 and the vector $(\mu_2, \dots, \mu_d) \rightsquigarrow \mathcal{N}(0, I_{d-1})$ whereas model \mathcal{M}_2 assumes that $(\mu_1, \dots, \mu_d) \rightsquigarrow \mathcal{N}(0, I_d)$. This simple example amounts at testing $\mu_1 = 0$ against $\mu_1 \neq 0$. Standard computations lead to $p(\mathcal{M}_1|\bar{\mathbf{x}} = (0, \dots, 0)) = \sqrt{M+1}/(1 + \sqrt{M+1})$. Similarly to the example of parameter estimation, the summary statistics $(\bar{\mathbf{x}}_2, \dots, \bar{\mathbf{x}}_M)$ convey no information and are added here to illustrate the curse of dimensionality. Figure 3 displays the mean squared error (MSE) of the two estimators of $p(\mathcal{M}_1|\bar{\mathbf{x}})$ using a total of $n = 10,000$ simulations and setting the percentage of accepted simulations to 5%. The local logistic regression provides a smaller MSE than the Nadaraya-Watson estimator for $d \geq 3$ and both estimators have comparable properties for $d = 1, 2$. The curse of dimensionality is once again displayed by this example since the MSEs increase with the dimension of the summary statistics. We note however that both estimators infer the probability of $p(\mathcal{M}_1|\bar{\mathbf{x}})$ accurately even when $d = 10$. Indeed the mean squared errors divided by the

square of the true value of $p(\mathcal{M}_1|\bar{\mathbf{x}})$ are equal to 0.65% (Nadaraya-Watson estimator) and 0.55% (local logistic regression).

Discussion

In this paper, we have presented Approximate Bayesian Computation as a technique of inference that relies on stochastic simulations and non-parametric statistics. We have introduced an estimator based on quadratic adjustment for which the bias involves less terms than the bias of the estimator with linear adjustment proposed by Beaumont et al. (2002). More generally, we have shown that the gain obtained with the estimators based on regression adjustment (equation (8) and (11)) is all the more important that the distribution of the residual ϵ in the model $\theta(\mathbf{s}) = m(\mathbf{s}) + \epsilon$ is independent of \mathbf{s} . This observation emphasizes the importance of model parameterization when considering estimators based on regression adjustment.

The crucial point raised by the asymptotic results given in Theorem 1 and 2 concerns the curse of dimensionality when performing Approximate Bayesian Computation. For both parameter estimation and model selection, Theorem 1 and 2 show that the properties of the ABC estimators may seriously deteriorate as the dimension of the summary statistics increases. This is a particularly acute issue for ABC since it is tempting to use as many summary statistics as possible so that not much information is lost when summarizing the data. To increase the probability of targeting close to the observed summary statistics and consequently improve the properties of ABC estimators, two types of alternative have been proposed. The first alternative

consists of reducing the dimensions of the summary statistics in the regression framework. Different techniques of dimension reduction have already been proposed in the context of ABC. Blum and François (2009) gave an estimator of the posterior distribution based on neural network regression and Lauenberger et al. (2009) proposed to reduce the number of summary statistics using principal component analysis or partial least-squares regression. The second type of alternative aims at performing simulations of the generative models in a parameter region for which the partial posterior distribution is substantial. Such adaptive ABC algorithm encompass ABC-MCMC algorithm (Marjoram et al 2003, Sisson et al 2007) and ABC sequential Monte Carlo sampler (Sisson et al. 2007; Beaumont et al. 2009; Toni et al. 2009).

Appendix: derivations of the main results

We will assume here the following conditions

- A1)** The kernel K has a finite second order moment such that $\int \mathbf{u}\mathbf{u}^T K(\mathbf{u}) d\mathbf{u} = \mu_2(K)I_d$ 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 $i = 1, \dots, d$. 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 partial evidence 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.

For the theorem 1, we assume that

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

For the theorem 2, we assume that

A5') The sequence of non-singular bandwidth matrices B is such that $1/(n|B|)$, and each entry of $B^t B$ tend to 0 as $n \rightarrow \infty$.

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 the variance of the Nadaraya-Watson estimator (see e.g. Pagan and Ullah 1999) comes from the fact that it is a ratio of two random variables. Following Pagan and Ullah (1999, page 98) or Scott (1992), we linearize the estimators in order to compute their biases and their variances. We write the estimators of the partial posterior distribution \hat{g}_j , $j = 0, 1, 2$, as

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

where

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

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

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

and

$$\hat{g}_D = \sum_{i=1}^n K_B(\mathbf{s}_i - \mathbf{s}_{obs}).$$

To compute the asymptotic expansions of the moments of the three estimators, we derive the following lemma (Pagan and Ullah 1999, page 98; Scott 1982)

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

$$\begin{aligned} \hat{g}_j(\theta|\mathbf{s}_{obs}) &= \frac{E[\hat{g}_{j,N}]}{E[\hat{g}_D]} + \frac{\hat{g}_{j,N} - E[\hat{g}_{j,N}]}{E[\hat{g}_D]} - \frac{E[\hat{g}_{j,N}](\hat{g}_D - E[\hat{g}_D])}{E[\hat{g}_D]^2} \\ &\quad + O_P(\text{Cov}(\hat{g}_{j,N}, \hat{g}_D) + \text{Var}[\hat{g}_D]) \end{aligned} \quad (23)$$

Proof. Lemma 1 is a simple consequence of the Taylor expansion for the function $(x, y) \mapsto x/y$ in the neighborhood of the point $(E[\hat{g}_{j,N}], E[\hat{g}_D])$ (see Pagan and Ullah 2004 for another proof). The order of the reminder follows from the weak law of large numbers. □

We now give an asymptotic expansion of all the expressions involved in equation (23).

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

$$E[\hat{g}_D] = p(\mathbf{s}_{obs}) + \frac{1}{2}\mu_2(K)\text{tr}(BB^t p_{ss}(\mathbf{s}_{obs})) + o(\text{tr}(B^t B)), \quad (24)$$

$$\begin{aligned} E[\hat{g}_{0,N}] &= p(\mathbf{s}_{obs})g(\theta|\mathbf{s}_{obs}) + \frac{1}{2}b'^2\mu_2(\tilde{K})g_{\theta\theta}(\theta|\mathbf{s}_{obs})p(\mathbf{s}_{obs}) \\ &+ \mu_2(K)[g_s(\theta|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}}^t BB^t p_s(\mathbf{s}_{obs}) + \frac{1}{2}g(\theta|\mathbf{s}_{obs})\text{tr}(BB^t p_{ss}(\mathbf{s}_{obs})) \\ &+ \frac{1}{2}p(\mathbf{s}_{obs})\text{tr}(BB^t g_{ss}(\theta|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}})] + o(b'^2) + o(\text{tr}(B^t B)), \end{aligned} \quad (25)$$

$$\begin{aligned} E[\hat{g}_{1,N}] &= p(\mathbf{s}_{obs})h(\epsilon|\mathbf{s}_{obs}) + \frac{1}{2}b'^2\mu_2(\tilde{K})h_{\epsilon\epsilon}(\epsilon|\mathbf{s}_{obs})p(\mathbf{s}_{obs}) \\ &+ \mu_2(K)[h_s(\epsilon|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}}^t BB^t p_s(\mathbf{s}_{obs}) + \frac{1}{2}h(\epsilon|\mathbf{s}_{obs})\text{tr}(BB^t p_{ss}(\mathbf{s}_{obs})) \\ &+ \frac{1}{2}p(\mathbf{s}_{obs})\text{tr}(BB^t h_{ss}(\epsilon|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}}) - \frac{h_{\epsilon}(\epsilon|\mathbf{s}_{obs})}{2}\text{tr}(BB^t m_{ss}(\mathbf{s}_{obs}))] \\ &+ o(b'^2) + o(\text{tr}(B^t B)), \end{aligned} \quad (26)$$

$$\begin{aligned} E[\hat{g}_{2,N}] &= p(\mathbf{s}_{obs})h(\epsilon|\mathbf{s}_{obs}) + \frac{1}{2}b'^2\mu_2(\tilde{K})h_{\epsilon\epsilon}(\epsilon|\mathbf{s}_{obs})p(\mathbf{s}_{obs}) \\ &+ \mu_2(K)[h_s(\epsilon|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}}^t BB^t p_s(\mathbf{s}_{obs}) + \frac{1}{2}h(\epsilon|\mathbf{s}_{obs})\text{tr}(BB^t p_{ss}(\mathbf{s}_{obs})) \\ &+ \frac{1}{2}p(\mathbf{s}_{obs})\text{tr}(BB^t h_{ss}(\epsilon|\mathbf{s})|_{\mathbf{s}=\mathbf{s}_{obs}}) + o(b'^2) + o(\text{tr}(B^t B)), \end{aligned} \quad (27)$$

$$\text{Var}[\hat{g}_D] = \frac{R(K)p(\mathbf{s}_{obs})}{n|B|} + O\left(\frac{1}{n}\right) + O\left(\frac{\text{tr}(BB^t)}{n|B|}\right), \quad (28)$$

$$\text{Var}[\hat{g}_{j,N}] = \frac{R(K)R(\tilde{K})g(\theta|\mathbf{s}_{obs})p(\mathbf{s}_{obs})}{nb'|B|} + O\left(\frac{1}{n}\right) + O\left(\frac{\text{tr}(BB^t)}{nb'|B|}\right) + O\left(\frac{b'}{n|B|}\right), \quad j = 0, 1, 2, \quad (29)$$

$$\text{Cov}[\hat{g}_{j,N}, \hat{g}_D] = \frac{R(K)p(\mathbf{s}_{obs})g(\theta|\mathbf{s}_{obs})}{n|B|} + O\left(\frac{1}{n}\right), \quad j = 0, 1, 2. \quad (30)$$

Proof.

Asymptotic expansion of $E[\hat{g}_D]$ By definition of \hat{g}_D , we have

$$E[\hat{g}_D] = E_{\mathbf{s}}[K_B(\mathbf{s} - \mathbf{s}_{obs})]$$

The computation of $E_{\mathbf{s}}[K_B(\mathbf{s} - \mathbf{s}_{obs})]$ is standard when computing the bias of multivariate density estimators and can be found for instance in Härdle et al. (2004, page 71). Here we describe the computations because the derivations of equations (25)-(30) are in the same vein. In the following we compute a Taylor expansion of $E_{\mathbf{s}}[K_B(\mathbf{s} - \mathbf{s}_{obs})]$. By definition of K_B , we have

$$E_{\mathbf{s}}[K_B(\mathbf{s} - \mathbf{s}_{obs})] = |B|^{-1} \int_{\mathbf{s}} K(B^{-1}(\mathbf{s} - \mathbf{s}_{obs}))p(\mathbf{s}) d\mathbf{s}.$$

Using the change of variable $\mathbf{u} = B^{-1}(\mathbf{s} - \mathbf{s}_{obs})$, we find that

$$E_{\mathbf{s}}[K_B(\mathbf{s} - \mathbf{s}_{obs})] = \int_{\mathbf{u}} K(\mathbf{u})p(B\mathbf{u} + \mathbf{s}_{obs}) d\mathbf{u}.$$

A Taylor expansion for p in the neighborhood of \mathbf{s}_{obs} gives

$$\begin{aligned} E_{\mathbf{s}}[K_B(\mathbf{s} - \mathbf{s}_{obs})] &= \int_{\mathbf{u}} K(\mathbf{u})\{p(\mathbf{s}_{obs}) + p_{\mathbf{s}}(\mathbf{s}_{obs})^t B\mathbf{u} \\ &\quad + \frac{1}{2}(B\mathbf{u})^t p_{\mathbf{ss}}(\mathbf{s}_{obs})(B\mathbf{u}) + o((B\mathbf{u})^t B\mathbf{u})\}. \end{aligned}$$

Using that $\int_{\mathbf{u}} K(\mathbf{u}) d\mathbf{u} = 1$ and $\int_{\mathbf{u}} \mathbf{u}_i K(\mathbf{u}) d\mathbf{u} = 0$, for $i = 1, \dots, d$, we find that

$$E_{\mathbf{s}}[K_B(\mathbf{s} - \mathbf{s}_{obs})] = p(\mathbf{s}_{obs}) + \int_{\mathbf{u}} K(\mathbf{u})\left\{\frac{1}{2}(B\mathbf{u})^t p_{\mathbf{ss}}(\mathbf{s}_{obs})(B\mathbf{u}) + o((B\mathbf{u})^t B\mathbf{u})\right\}.$$

The second term on the right-hand side of the previous equation can be found using a standard result for the expectation of a quadratic form. This result states that $E[\mathbf{u}^t \Lambda \mathbf{u}] = \text{tr}(\Lambda \Sigma) + \mu^t \Lambda \mu$ where μ and Σ are the expectation and the covariance matrix of \mathbf{u} . As a consequence, we have

$$E_{\mathbf{s}}[K_H(\mathbf{s} - \mathbf{s}_{obs})] = p(\mathbf{s}_{obs}) + \text{tr}(B^t p_{\mathbf{ss}}(\mathbf{s}_{obs})B) + o(\text{tr}(B^t B)).$$

Using the cyclic property of the trace, we find the asymptotic expansion of $E[\hat{g}_D]$ given in equation (24).

Asymptotic expansion of $E[\hat{g}_{0,N}]$ By definition of $\hat{g}_{1,N}$, we have

$$E[\hat{g}_{0,N}] = E_{\theta_0, \mathbf{s}}[\tilde{K}_{b'}(\theta_0 - \theta)K_B(\mathbf{s} - \mathbf{s}_{obs})].$$

We have

$$E_{\theta_0, \mathbf{s}}[\tilde{K}_{b'}(\theta_0 - \theta)K_B(\mathbf{s} - \mathbf{s}_{obs})] = \int_{\mathbf{s}, \theta_0} \tilde{K}_{b'}(\theta_0 - \theta)K_B(\mathbf{s} - \mathbf{s}_{obs})p(\mathbf{s})g(\theta_0|\mathbf{s}) d\theta_0 d\mathbf{s}.$$

Using the change of variable $\mathbf{u} = B^{-1}(\mathbf{s} - \mathbf{s}_{obs})$, we have

$$E_{\theta_0, \mathbf{s}}[\tilde{K}_{b'}(\theta_0 - \theta)K_B(\mathbf{s} - \mathbf{s}_{obs})] = \int_{\mathbf{u}, \theta_0} \tilde{K}_{b'}(\theta_0 - \theta)K(\mathbf{u})p(\mathbf{s}_{obs} + B\mathbf{u})g(\theta_0|\mathbf{s}_{obs} + B\mathbf{u}) d\theta_0 d\mathbf{u}$$

Using a Taylor expansion in the neighborhood of \mathbf{s}_{obs} and the fact that the terms of order 1 vanish, we find

$$\begin{aligned} E_{\theta_0, \mathbf{s}}[\tilde{K}_{b'}(\theta_0 - \theta)K_B(\mathbf{s} - \mathbf{s}_{obs})] &= \int_{\mathbf{u}, \theta_0} \tilde{K}_{b'}(\theta_0 - \theta)K(\mathbf{u})\{p(\mathbf{s}_{obs})g(\theta_0|\mathbf{s}_{obs}) \\ &+ (B\mathbf{u})^t[p_{\mathbf{s}}(\mathbf{s}_{obs})g_{\mathbf{s}}(\theta_0|\mathbf{s})]_{\mathbf{s}=\mathbf{s}_{obs}} + \frac{1}{2}p_{\mathbf{ss}}(\mathbf{s}_{obs})g(\theta_0|\mathbf{s}_{obs}) \\ &+ \frac{1}{2}p(\mathbf{s}_{obs})g_{\mathbf{ss}}(\theta_0|\mathbf{s})_{\mathbf{s}=\mathbf{s}_{obs}} + o(1)](B\mathbf{u})\} \end{aligned}$$

Introducing the change of variable $\tau = (\theta_0 - \theta)/h'$ and ignoring the terms of degree 1 in h' that vanish, we get equation (25).

Asymptotic expansion of $E[\hat{g}_{1,N}]$ We start by introducing some notations. First recall that $h(\cdot|\mathbf{s})$ denotes the conditional distribution of the residual $\epsilon_0 = \theta_0 - m(\mathbf{s})$ so that $g(\epsilon_0 + m(\mathbf{s})|\mathbf{s}) = h(\epsilon_0|\mathbf{s})$. We introduce

the error function $d1$ such that $d1(\mathbf{s}) = \hat{m}_1(\mathbf{s}) - m(\mathbf{s})$. In the following, we will make use of the following expressions that follow by application of the derivation rules

$$d1_{\mathbf{s}}(\mathbf{s}) = \hat{\beta} - m_{\mathbf{s}}(\mathbf{s}), \quad (31)$$

and

$$d1_{\mathbf{ss}}(\mathbf{s}) = -m_{\mathbf{ss}}(\mathbf{s}). \quad (32)$$

We start by writing

$$E[\hat{g}_{1,N}] = E_{\theta_0, \mathbf{s}, \hat{m}_1}[\tilde{K}_{b'}(\theta_0^* - \theta)K_B(\mathbf{s} - \mathbf{s}_{obs})],$$

where $\theta_0^* = \hat{m}_1(\mathbf{s}_{obs}) + (\theta_0 - \hat{m}_1(\mathbf{s}))$. This can be rewritten as

$$\begin{aligned} E[\hat{g}_{1,N}] &= E_{\theta_0, \mathbf{s}, \hat{m}_1}[\tilde{K}_{b'}((\theta_0 - m(\mathbf{s})) - (\theta_0 - m(\mathbf{s}_{obs})) + (d1(\mathbf{s}_{obs}) - d1(\mathbf{s})))K_B(\mathbf{s} - \mathbf{s}_{obs})] \\ &= E_{\epsilon_0, \mathbf{s}, \hat{m}_1}[\tilde{K}_{b'}(\epsilon_0 - \epsilon + (d1(\mathbf{s}_{obs}) - d1(\mathbf{s})))K_B(\mathbf{s} - \mathbf{s}_{obs})], \end{aligned} \quad (33)$$

where $\epsilon = \theta_0 - m(\mathbf{s}_{obs})$ and $\epsilon_0 = \theta_0 - m(\mathbf{s})$. We introduce the change of variable $\tau = \frac{\epsilon_0 - \epsilon + (d1(\mathbf{s}_{obs}) - d1(\mathbf{s}))}{b'}$, so that

$$E[\hat{g}_{1,N}] = \int_{\tau, \mathbf{s}} E_{\hat{m}_1|\mathbf{s}, \tau}[\tilde{K}(\tau)K_B(\mathbf{s} - \mathbf{s}_{obs})p(\mathbf{s})h(\epsilon + b'\tau + d1(\mathbf{s}) - d1(\mathbf{s}_{obs})|\mathbf{s})].$$

The next step consists of considering the now classic change of variable $\mathbf{u} = B^{-1}(\mathbf{s} - \mathbf{s}_{obs})$ and performing a Taylor expansion in the neighborhood of \mathbf{s}_{obs} . The Taylor expansion relies on equations (31) and (32) that give the derivatives of the function $d1$. After, a computation in which the terms of degree 1 in \mathbf{u} and τ are neglected because they vanish, we find the formula (26). Note that the result relies on the fact that $\hat{\beta}$ is a consistent estimator of $m_{\mathbf{s}}(\mathbf{s}_{obs})$ (Ruppert and Wand 1994).

Asymptotic expansion of $E[\hat{g}_{2,N}]$ The derivation of $E[\hat{g}_{2,N}]$ is similar to the derivation of $E[\hat{g}_{1,N}]$ except that we now introduce the function

$$d2(\mathbf{s}) = \hat{m}_2(\mathbf{s}) - m(\mathbf{s}), \quad \mathbf{s} \in \mathcal{R}^d.$$

By following the same line of proofs as before and using that $\hat{\gamma}$ is a consistent estimator of $m_{ss}(\mathbf{s}_{obs})$, we find equation (27).

Asymptotic expansion of $Var[\hat{g}_D]$ The computation of $Var[\hat{g}_D]$ can be found in Härdle et al. (2004, page 71). Since it is a simple computation, we describe briefly the computation. By definition of \hat{g}_D , we have

$$Var[\hat{g}_D] = \frac{1}{n} \int_{\mathbf{s}} K_B(\mathbf{s} - \mathbf{s}_{obs})^2 p(\mathbf{s}) d\mathbf{s} - \frac{1}{n} E[\hat{g}_D]^2$$

Using the the change of variable $u = B^{-1}(\mathbf{s} - \mathbf{s}_{obs})$, we find

$$Var[\hat{g}_D] = \frac{1}{n|B|} \int_{\mathbf{u}} K(\mathbf{u})^2 p(\mathbf{s}_{obs} + B\mathbf{u}) + O\left(\frac{1}{n}\right)$$

By using a Taylor expansion in the neighborhood \mathbf{s}_{obs} , we find equation (28).

Asymptotic expansion of $Var[\hat{g}_{0,N}]$ We have

$$Var[\hat{g}_{0,N}] = \frac{1}{n} \int_{s, \theta_0} \tilde{K}_{b'}(\theta_0 - \theta)^2 K_B(\mathbf{s} - \mathbf{s}_{obs})^2 g(\theta_0|\mathbf{s}) f(\mathbf{s}) ds d\theta_0 - \frac{1}{n} E[\hat{g}_{0,N}]^2$$

The two standard changes of variables gives

$$Var[\hat{g}_{0,N}] = \frac{1}{nb'|B|} \int_{\mathbf{u}, \tau} \tilde{K}(\tau)^2 K(\mathbf{u})^2 g(\theta_0 + \tau b'|\mathbf{s}_{obs} + B\mathbf{u}) p(\mathbf{s}_{obs} + B\mathbf{u}) d\mathbf{u} d\tau + O\left(\frac{1}{n}\right).$$

A Taylor expansion now gives the result of equation (29).

Asymptotic expansion of $\text{Var}[\hat{g}_{1,N}]$ and $\text{Var}[\hat{g}_{2,N}]$ Both computations are similar and we will restrict our analysis to the computation of $\text{Var}[\hat{g}_{1,N}]$. The details of the computations will be omitted here. Up to a constant $\hat{g}_{1,N}$ is a sum of random variables. By contrast to the computation of $\text{Var}[\hat{g}_{0,N}]$, the random variables are not independent anymore and a term accounting for the covariance should be included. Using a Taylor expansion, it can be shown that the term corresponding to the covariance is negligible compared to the term that accounts for the sum of variances. This consideration leads to

$$\text{Var}[\hat{g}_{1,N}] = \frac{1}{n} \int_{\mathbf{s}, \theta_0} E_{\hat{m}_1|\mathbf{s}, \theta_0} [\tilde{K}_{b'}(\theta_0 - m(\mathbf{s}) - (\theta - m(\mathbf{s}_{obs})) + d1(\mathbf{s}_{obs}) - d1(\mathbf{s}))^2 K_B(\mathbf{s} - \mathbf{s}_{obs})^2 g(\theta_0|\mathbf{s}) p(\mathbf{s})] d\mathbf{s} d\theta_0 + O\left(\frac{1}{n}\right)$$

Using the two changes of variables $\mathbf{u} = B^{-1}(\mathbf{s} - \mathbf{s}_{obs})$, $\tau = \frac{\theta_0 - m(\mathbf{s}) - (\theta - m(\mathbf{s}_{obs})) + (d1(\mathbf{s}_{obs}) - d1(\mathbf{s}))}{b'}$, and a Taylor expansion, we find

$$\text{Var}[\hat{g}_{1,N}] = \frac{R(K)R(\tilde{K})p(\mathbf{s}_{obs})h(\epsilon|\mathbf{s}_{obs})}{n|B|b'} + O\left(\frac{1}{n}\right) + O\left(\frac{\text{tr}(BB^t)}{nb'|B|}\right) + O\left(\frac{b'}{n|B|}\right).$$

By using that $h(\epsilon|\mathbf{s}_{obs}) = g(\theta|\mathbf{s}_{obs})$, we get the result given in equation (29).

Asymptotic expansion of $\text{Cov}[\hat{g}_D, \hat{g}_{j,N}]$ We have

$$\text{Cov}[\hat{g}_D, \hat{g}_{0,N}] = \frac{1}{n^2} \sum_{i \neq j} E[K_B(\mathbf{s}_i - \mathbf{s}_{obs}) \tilde{K}_{b'}(\theta_j - \theta_0) K_B(\mathbf{s}_j - \mathbf{s}_{obs})] - E[\hat{g}_D] E[\hat{g}_{0,N}].$$

This leads to

$$\begin{aligned} \text{Cov}[\hat{g}_D, \hat{g}_{0,N}] &= \frac{n}{n^2} \int_{\mathbf{s}, \theta_0} K_B(\mathbf{s} - \mathbf{s}_{obs})^2 \tilde{K}_{b'}(\theta - \theta_0) g(\theta_0|\mathbf{s}) p(\mathbf{s}) d\theta_0 d\mathbf{s} \\ &\quad + E_{\mathbf{s}}[K_B(\mathbf{s} - \mathbf{s}_{obs})] E_{\mathbf{s}, \theta_0}[K_B(\mathbf{s} - \mathbf{s}_{obs}) \tilde{K}_{b'}(\theta - \theta_0)] \left(\frac{n(n-1)}{n^2} - 1\right) \end{aligned}$$

Because the last term on the right hand side is of the order of $1/n$, we get

$$\text{Cov}[\hat{g}_D, \hat{g}_{0,N}] = \frac{R(K)p(\mathbf{s}_{obs})g(\theta|\mathbf{s}_{obs})}{n|B|} + O\left(\frac{1}{n}\right). \quad (34)$$

The computations of $\text{Cov}[\hat{g}_D, \hat{g}_{j,N}]$, for $j = 1, 2$, are similar and are omitted here.

□

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 3 *Assume that B is a non-singular bandwidth matrix and assume that conditions (A1)-(A5) holds, then the bias of \hat{g}_j , $j = 0, 1, 2$, is given by*

$$E[\hat{g}_j(\theta|\mathbf{s}_{obs}) - g(\theta|\mathbf{s}_{obs})] = D_1 b'^2 + D_{2,j} + O_P((\text{tr}(B^t B) + b')^2) + O_P\left(\frac{1}{n|B|}\right), \quad j = 0, 1, 2, \quad (35)$$

with

$$\begin{aligned} D_1 &= C_1 = \frac{\mu_2(\tilde{K})g_{\theta\theta}(\theta|\mathbf{s}_{obs})}{2}, \\ D_{2,0} &= \mu_2(K) \left(\frac{g_{\mathbf{s}}(\theta|\mathbf{s})^t_{|\mathbf{s}=\mathbf{s}_{obs}} BB^t p_{\mathbf{s}}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})} + \frac{\text{tr}(BB^t g_{\mathbf{s}\mathbf{s}}(\theta|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}})}{2} \right), \\ D_{2,1} &= \mu_2(K) \left(\frac{h_{\mathbf{s}}(\epsilon|\mathbf{s})^t_{|\mathbf{s}=\mathbf{s}_{obs}} BB^t p_{\mathbf{s}}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})} + \frac{\text{tr}(BB^t h_{\mathbf{s}\mathbf{s}}(\epsilon|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}})}{2} - \frac{h_{\epsilon}(\epsilon|\mathbf{s}_{obs})\text{tr}(BB^t m_{\mathbf{s}\mathbf{s}})}{2} \right), \end{aligned}$$

and

$$D_{2,2} = \mu_2(K) \left(\frac{h_{\mathbf{s}}(\epsilon|\mathbf{s})^t_{|\mathbf{s}=\mathbf{s}_{obs}} BB^t p_{\mathbf{s}}(\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})} + \frac{\text{tr}(BB^t h_{\mathbf{s}\mathbf{s}}(\epsilon|\mathbf{s})_{|\mathbf{s}=\mathbf{s}_{obs}})}{2} \right),$$

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

$$\text{Var}[\hat{g}_j(\theta|\mathbf{s}_{obs})] = \frac{R(K)R(\tilde{K})g(\theta|\mathbf{s}_{obs})}{p(\mathbf{s}_{obs})n|B|b'}(1 + o_P(1)), \quad j = 0, 1, 2, \quad (36)$$

Proof.

Theorem 3 is a consequence of Lemma 1 and 2. Taking expectations on both sides of equation (23), we find that

$$E[\hat{g}_j(\theta|\mathbf{s}_{obs})] = \frac{E[\hat{g}_{j,N}]}{E[\hat{g}_D]} + O_P[\text{Cov}(g_{j,N}, \hat{g}_D) + \text{Var}(\hat{g}_D)]. \quad (37)$$

Using a Taylor expansion, and the equations (24)-(27), (28), and (30) given in Lemma 2, we find the bias of the estimators given in equation (35).

For the computation of the variance, we find from equation (23) and (37) that

$$\hat{g}_j(\theta|\mathbf{s}_{obs}) - E[\hat{g}_j(\theta|\mathbf{s}_{obs})] = \frac{\hat{g}_{j,N} - E[\hat{g}_{j,N}]}{E[\hat{g}_D]} - \frac{E[\hat{g}_{j,N}](\hat{g}_D - E[\hat{g}_D])}{E[\hat{g}_D]^2} + O_P\left(\frac{1}{n|B|}\right). \quad (38)$$

The order of the reminder follows from equations (28) and (30). Taking the expectation of the square of equation (38), we now find

$$\text{Var}[\hat{g}_j(\theta|\mathbf{s}_{obs})] = \frac{\text{Var}[\hat{g}_{j,N}]}{E[\hat{g}_D]} + \frac{E[\hat{g}_{j,N}]^2 \text{Var}[\hat{g}_D]}{E[\hat{g}_D]^4} - 2\text{Cov}(\hat{g}_D, \hat{g}_{j,N}) \frac{E[\hat{g}_{j,N}]}{E[\hat{g}_D]^3} + o_P\left(\frac{1}{n|B|b'}\right). \quad (39)$$

The variance of the estimators given in equation (36) follows from a Taylor expansion that makes use of equations (24)-(30) given in Lemma 2.

□

References

- [1] Abril JC (1994) On the concept of approximate sufficiency. *Pakistan J Stat* 10: 171–177
- [2] Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics* 162: 2025–2035
- [3] Beaumont MA (2008) Joint determination of topology, divergence time, and immigration in population trees. In Matsumura S, Forster P, Renfrew C. *Simulation, Genetics and Human Prehistory*, pp. 134154. McDonald Institute Monographs: Cambridge McDonald Institute for Archeological Research, UK
- [4] Beaumont MA, Cornuet J-M, Marin J-M, Robert CP (2009) Adaptivity for ABC algorithms: the ABC-PMC scheme. *arXiv:0805.2256*
- [5] Blum MGB, Tran VC (2009) HIV with contact-tracing: a case study in Approximate Bayesian Computation. *ArXiv:0810.0896*
- [6] Blum MGB, O Franois (2009) Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing*, in press
- [7] Bortot P, Coles SG and Sisson SA (2007) Inference for stereological extremes. *J Amer Statistical Assoc* 102: 84–92
- [8] Box GEP, Cox DR (1964) An analysis of transformations. *J Roy Statist Soc B* 26: 211–246

- [9] Cabrera J, Yohai VJ (1999) A New Computational Approach for Bayesian and Robust Bayesian Statistical Analysis, www.rci.rutgers.edu/cabrera/pap/vic.pdf, preprint
- [10] De Gooijer JG, Zerom D (2003) On conditional density estimation. *Stat Neerl* 57: 159-176
- [11] Diggle PJ, Gratton RJ (1984) Monte Carlo methods of inference for implicit statistical models. *J Roy Statist Soc B* 46: 193-227
- [12] Doksum KA, Lo AY (1990) Consistent and robust Bayes procedures for location based on partial information. *Annals of Statistics* 18: 443-453
- [13] Fan J (1992) Design-adaptive nonparametric regression. *J Amer Statist Assoc* 87: 998-1004
- [14] Fan J, NE Heckman, MP Wand (1995) Local polynomial kernel regression for generalized linear models and quasi-likelihood functions *J Amer Statist Assoc* 90: 141-150
- [15] Fan J, Yao Q, Tong H (1996) Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* 83: 189-196
- [16] Fan J, Q Yao (1998) Efficient Estimation of Conditional Variance Functions in Stochastic Regression. *Biometrika* 85: 645-660
- [17] Fan J, Yim TH (2004) A data-driven method for estimating conditional densities. *Biometrika* 91: 819-834

- [18] Fagundes NJR, Ray N, Beaumont M, Neuenschwander S, Salzano SM, Bonatto SL, Excoffier L (2007) Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci USA* 104: 17614-17619
- [19] François O, Blum MGB, Jakobsson M, Rosenberg NA (2008) Demographic history of European populations of *Arabidopsis thaliana*. *PLoS Genetics*, 4(5): e1000075
- [20] Fu Y-X, Li W-H (1997) Estimating the age of the common ancestor of a sample of DNA sequences. *Mol Biol Evol* 14: 195–199
- [21] Gourieroux C, Monfort A, Renault E (1993) Indirect Inference. *J Appl Econom* 8: 85–118
- [22] Grelaud A, Robert CP, Marin J-M, Rodolphe F, Taly J-F (2008) ABC methods for model choice in Gibbs random fields. *arXiv:0807.2767*
- [23] Hall P, Wolff RCL, Yao Q (1999) Methods for estimating a conditional distribution function. *J. Amer. Statist. Assoc.* 94: 154–163
- [24] Hall P, Racine J, Li Q (2004) Cross-validation and the estimation of conditional probability densities. *J. Amer. Statist. Assoc.* 99: 1015–1026
- [25] Hansen B (2004) Nonparametric Conditional Density Estimation. <http://www.ssc.wisc.edu/~bhansen/papers/ncde.pdf>, preprint
- [26] Härdle W, M Müller, Sperlich S, Werwatz A (2004) Nonparametric and Semiparametric Models. Springer-Verlag.

- [27] Hyndman RJ, Bashtannyk DM, Grunwald GK (1996) Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics* 5: 315-336
- [28] Jabot F, Chave J (2009) Inferring the parameters of the neutral theory of biodiversity using phylogenetic information and implications for tropical forests. *Ecol Lett* 12: 239-248
- [29] Joyce P, Marjoram P (2008) Approximately Sufficient Statistics and Bayesian Computation. *Statistical Applications in Genetics and Molecular Biology* 7: 26
- [30] Le Cam L. (1964) Sufficiency and approximate sufficiency. *Ann. Math. Statist.* 35: 1419–1455
- [31] Leuenberger C, Wegmann D, Excoffier L (2009) Bayesian Computation and Model Selection in Population Genetics. *arXiv:0901.2231*
- [32] Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci USA* 100: 15324–15328
- [33] McCullagh P, JA Nelder (1989) *Generalized Linear Models*. Chapman and Hall: London.
- [34] Nadaraya, E A (1964) On Estimating Regression. *Theory of Probability and its Applications* 9: 141-142
- [35] Pagan A, Ullah A (1999) *Nonparametric Econometrics*. Cambridge University Press: Cambridge, UK.

- [36] Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* 16: 1791–1798
- [37] R Development Core Team (2008) R: A Language and Environment for Statistical. R Foundation for Statistical Computing, Vienna, Austria
- [38] Ratmann O, Jørgensen O, Hinkley T, Stumpf M, Richardson S, Wiuf C (2007) Using likelihood-free inference to compare evolutionary dynamics of the protein networks of *H. pylori* and *P. falciparum*. *PLoS Comp Biol*, 3:e230
- [39] Rosenblatt M (1969) Conditional probability density and regression estimates in Multivariate Analysis II, Ed. P.R. Krishnaiah, pp. 25-31. New York: Academic Press
- [40] Ruppert D, Wand MP (1994) Multivariate locally weighted least squares regression. *Annals of Statistics* 22: 1346-1370
- [41] Scott DW (1992) Multivariate Density Estimation: Theory, Practice, and Visualization. Wiley, New York
- [42] Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall: London
- [43] Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. *Proc Natl Acad Sci USA* 104: 1760–1765

- [44] Tanaka M, Francis A, Luciani F, Sisson S (2006) Estimating tuberculosis transmission parameters from genotype data using approximate Bayesian computation. *Genetics* 173: 1511–1520
- [45] Tavaré S, Balding DJ, Griffiths RC, Donnelly P (1997) Inferring coalescence times from DNA sequence data. *Genetics* 145: 505–518
- [46] Toni T, Welch D, Strelkowa N, Stumpf MPH (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface* 6: 187–202
- [47] Watson GS (1964) Smooth regression analysis. *Shankya Series A* 26: 359–372
- [48] Weiss G, von Haeseler A (1998) Inference of population history using a likelihood approach. *Genetics* 149: 1539–1546
- [49] Wilkinson RD (2009). Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. *arXiv:0811.3355*
- [50] Wilkinson RD, Tavaré S (2009). Estimating the primate divergence time using conditioned birth-and-death processes. *Theor Popul Biol.* (2009), in press

FIGURE LEGENDS

Figure 1. Minimum number of simulations that are required to reach a given level of accuracy as a function of the dimension of the summary statistics. The minimum number of simulations is defined as the smallest number of simulations that are required so that the relative mean square error is less than 10%.

Figure 2. Mean integrated square error as a function of the dimension of the summary statistics. The parameter to infer is the exponential of the location parameter e^{μ_1} of the first component of a gaussian sample. The total number of simulations was set to $n = 10,000$ and the percentage of accepted simulation was set to 5%.

Figure 2. Mean square error of the probability of model \mathcal{M}_1 as a function of the dimension of the summary statistics. The total number of simulations was set to $n = 10,000$ and the percentage of accepted simulation was set to 5%.

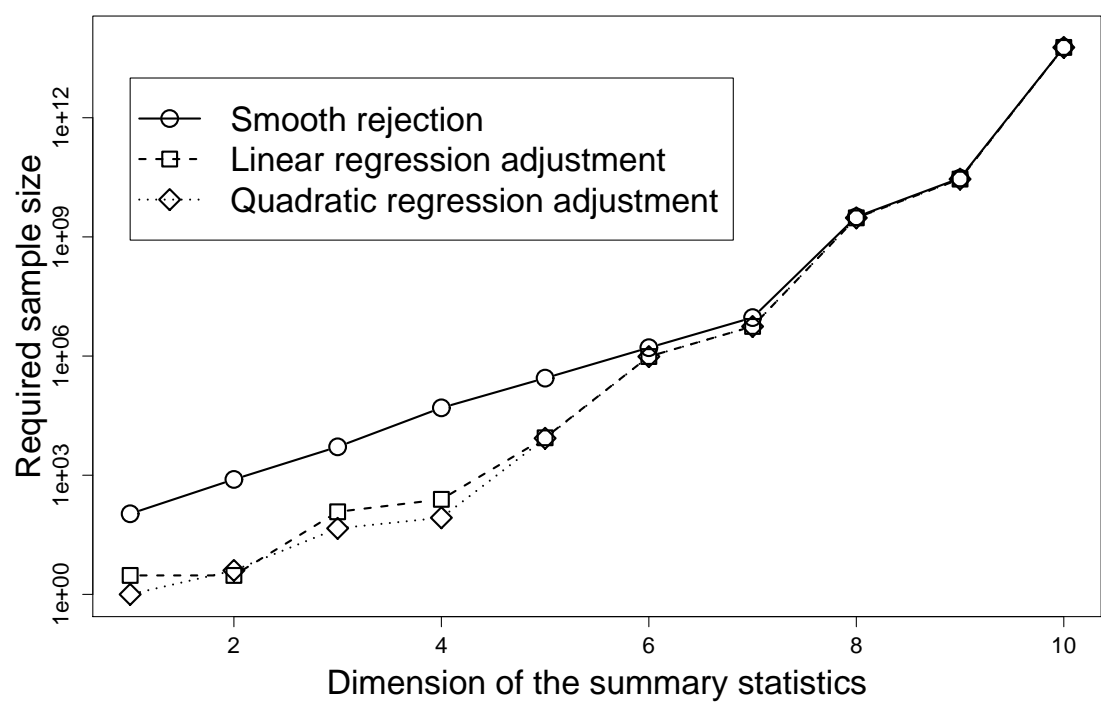


Figure 1:

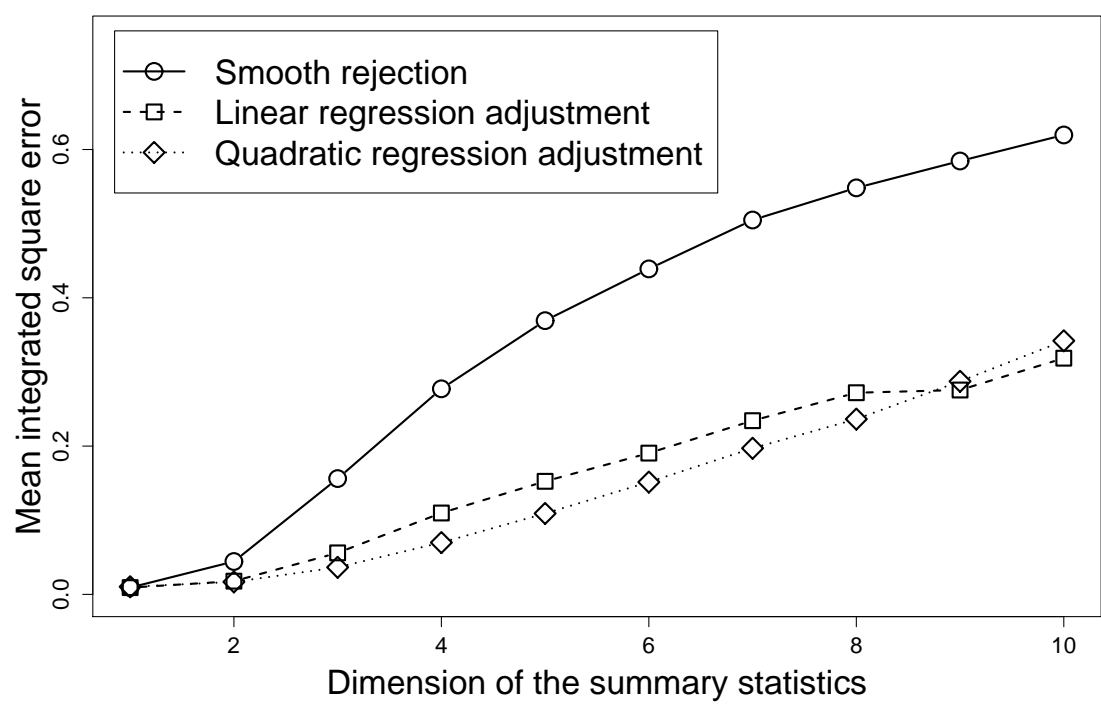


Figure 2:

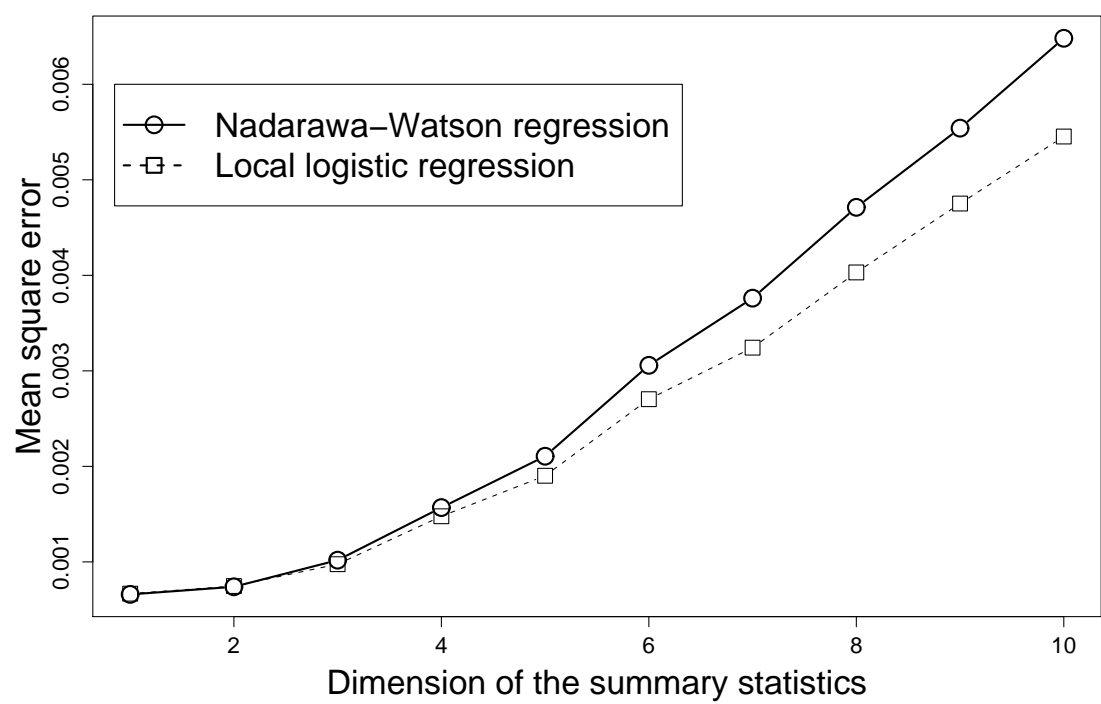


Figure 3: