Weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and implications in ABC

Samuel Soubeyrand, Emilie Haon-Lasportes

Weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and implications in ABC

Samuel Soubeyrand[1] and Emilie Haon–Lasportes

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

Abstract

The weak convergence of posterior distributions conditional on maximum pseudo-likelihood estimates (MPLE) is studied and exploited to justify the use of MPLE as summary statistics in approximate Bayesian computation (ABC). Our study could be generalized by replacing the pseudo-likelihood by other estimating functions (e.g. quasi-likelihoods and contrasts).

Keywords: Approximate Bayesian computation, Bernstein - von Mises theorem, Weak convergence.

2010 MSC: 62F12, 62Fxx, 65C60

1. Introduction

Approximate Bayesian computation (ABC) has been developed to make Bayesian inference with models that can be used to generate data sets but whose probability distribution of state variables is intractable (Marin et al., 2012). The intractability of this distribution makes impossible the application of the exact Bayesian approach, even by using numerical algorithms. ABC provides a sample from the parameter space that is approximately distributed under a posterior distribution of parameters conditional on sum-

[1]Corresponding Author: Samuel.Soubeyrand@avignon.inra.fr

mary statistics. In general, this posterior does not coincide with the posterior distribution of parameters conditional on full raw data. Here, we are interested in the specific case where (some of) the summary statistics are point estimates of parameters (PEP), as in Drovandi et al. (2011), Fearnhead and Prangle (2012), Gleim and Pigorsch (2013) and Mengersen et al. (2013).

In the classical Bayesian framework, posteriors conditional on PEP can be viewed as specific cases of posteriors conditional on partial information (Doksum and Lo, 1990; Soubeyrand et al., 2009). Here, we provide new results of weak convergence when PEP are either maximum likelihood estimates (MLE) or pseudo-maximum likelihood estimates (MPLE). The case where PEP are MPLE is of specific interest because, in ABC, it may be possible to compute MPLE via simplifications of the dependence structure in the model, and to use MPLE as summary statistics.

The results of weak convergence that are provided in this note can be viewed as new extensions of the Bernstein – von Mises (BvM) theorem. For parametric models from which independent observations are made, the BvM theorem (i) states conditions under which the posterior distribution is asymptotically normal and (ii) subsequently leads to the efficiency of Bayesian point estimators and to the convergence of Bayesian confidence sets to frequentist limit confidence sets (Walker, 1969; Freedman, 1999). Thus, the BvM theorem can be viewed as a frequentist justification of posterior distributions for the estimation of parameters. Numerous extensions of the BvM theorem have been proposed, for instance, when the model is semiparametric or nonparametric (Bickel and Kleijn, 2012; Bontemps, 2011; Castillo, 2012a,b; Castillo and Nickl, 2013; Rivoirard and Rousseau, 2012), when observations are dependent (Borwanker et al., 1971; Tamaki, 2008),

when the model is misspecified (Kleijn and van der Vaart, 2012), and when the model is nonregular (e.g. when the true value of the parameter is on the boundary of the parameter space; Bochkina and Green, 2014).

Here, we extend the BvM theorem (i) when raw observations are replaced by the MLE (Lemma 1) or an MPLE (Lemma 2), and (ii) when the posterior conditional on an MPLE is approximated via ABC (Theorem 1). Using a posterior distribution (approximate or not) conditional on an MPLE, that was built by ignoring some dependences in the model, can be viewed as using a misspecified model like in Kleijn and van der Vaart (2012).

The BvM extensions obtained in the classical Bayesian framework (Point (i) in the paragraph above) are viewed as stepping stones that lead to the BvM extension obtained in the ABC framework (Point (ii)). Advancing theory in ABC has generally no direct practical implications because assumptions that may be required to prove theorems cannot be checked for a real-life implicit stochastic model whose distribution theory is intractable. However, showing an analytic result for a large class of theoretically tractable models may lead to conjecture that the result holds for some stochastic implicit models. Specifically, the work presented here allows us to conjecture that (i) an ABC–posterior distribution conditional on an MPLE is asymptotically normal and centered around the MPLE, and (ii) resulting point estimates and confidence sets converge to their frequentist analogues.

2. Main notations

Observed data $\mathcal{D} \in \mathbb{D}$ are assumed to be generated under the stochastic model $\mathcal{M}_\theta$ parametrized by $\theta \in \Theta$ with prior density π. The data space $\mathbb{D}$ and the parameter space $\Theta \subset \mathbb{R}^q$ ($q \in \mathbb{N}^*$) are both included in

multidimensional sets of real vectors. The probability distribution functions (p.d.f.) of the model and the prior are defined with respect to the Lebesgue measure. Let $p(\mathcal{D} \mid \theta)$ denote the likelihood of the model and $p(\theta \mid \mathcal{D}) = p(\mathcal{D} \mid \theta)\pi(\theta)/p(\mathcal{D})$ the full sample posterior of the parameter vector θ. The vector $\hat{\theta}_{ML} \in \Theta$ is the maximum likelihood estimate (MLE) of θ: $\hat{\theta}_{ML} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ p(\mathcal{D} \mid \theta)$. The posterior of parameters conditional on the MLE is $p(\theta \mid \hat{\theta}_{ML}) = p(\hat{\theta}_{ML} \mid \theta)\pi(\theta)/p(\hat{\theta}_{ML})$, where $p(\hat{\theta}_{ML} \mid \theta)$ is the p.d.f. of the MLE given θ. Besides, we are interested in models whose likelihoods are not tractable because of the dependence structure in the data, but for which we can build tractable pseudo-likelihoods, say $\tilde{p}(\mathcal{D} \mid \theta)$. A pseudo-likelihood is generally built by ignoring some of the dependencies in the data (Gaetan and Guyon, 2008; Gourieroux et al., 1983). The vector $\hat{\theta}_{MPL} \in \Theta$ is a maximum pseudo-likelihood estimate (MPLE) of θ: $\hat{\theta}_{MPL} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \tilde{p}(\mathcal{D} \mid \theta)$. The posterior of parameters conditional on the MPLE is $p(\theta \mid \hat{\theta}_{MPL}) = p(\hat{\theta}_{MPL} \mid \theta)\pi(\theta)/p(\hat{\theta}_{MPL})$, where $p(\hat{\theta}_{MPL} \mid \theta)$ is the p.d.f. of the MPLE given θ.

3. Posterior conditional on the MLE

The full sample posterior $p(\theta \mid \mathcal{D})$ and the posterior conditional on the MLE $p(\theta \mid \hat{\theta}_{ML})$ exactly coincide in specific cases (e.g. when the MLE are sufficient statistics), but do not coincide in general. Our aim, in this section, is to provide an asymptotically equivalent distribution for $p(\theta \mid \hat{\theta}_{ML})$.

Bernstein – von Mises (BvM) theorems provide, for various statistical models, the asymptotic behavior of posteriors distributions. For example, following Walker (1969) and Lindley (1965, p. 130), we consider a set $\mathcal{D} = (D_1, \ldots, D_n)$ of n i.i.d. variables drawn from a parametric distribution with

density $f(\cdot \mid \theta)$ with respect to a σ-finite measure on the real line, where θ is in $\Theta \subset \mathbb{R}^q$. Under this setting and additional regularity conditions, the BvM theorem establishes the asymptotic normality of the full sample posterior (Walker, 1969, Theorem 2 and conclusion): the full sample posterior density of θ is, for large n, equivalent to the normal density with mean vector equal to the MLE $\hat{\theta}_{ML}$ and covariance matrix equal to $\Omega_n(\hat{\theta}_{ML})^{-1}$:

$$p(\theta \mid \mathcal{D}) \underset{n\to\infty}{\sim} \phi_{\hat{\theta}_{ML},\Omega_n(\hat{\theta}_{ML})^{-1}}(\theta),$$

where $\phi_{\mu,\Sigma}$ denotes the density of the normal distribution with mean vector μ and covariance matrix Σ, and $\Omega_n(\alpha)$ is the $q \times q$ matrix with element (i,j) equal to $\left(-\partial^2 \log p(\mathcal{D} \mid \theta)/\partial\theta_i \partial\theta_j\right)_{\theta=\alpha}$.

To provide an asymptotically equivalent distribution for $p(\theta \mid \hat{\theta}_{ML})$ as in BvM theorems, we assume in Lemma 1 (see below) that the MLE is asymptotically normal and consistent. For example, consider the same statistical model than above and assume that assumptions made in Lehmann and Casella (1998, Theorem 5.1 of the MLE asymptotic normality, p. 463) are satisfied. In particular, assume that data were generated with parameter vector θ. Then, the density of $\hat{\theta}_{ML}$ is, for large n and given θ, equivalent to the normal density with mean vector equal to the true parameter vector θ and covariance matrix equal to $n^{-1}I(\theta)^{-1}$:

$$p(\hat{\theta}_{ML} \mid \theta) \underset{n\to\infty}{\sim} \phi_{\theta,n^{-1}I(\theta)^{-1}}(\hat{\theta}_{ML}). \tag{1}$$

where $I(\theta)$ denotes the $q \times q$ Fisher information matrix.

Lemma 1 (Asymptotic normality of the posterior conditional on the MLE)**.** *Consider the modeling setting described in Section 2 and suppose that the MLE satisfies Equation* (1) *with non-singular matrix* $I(\theta)$*. Assume in addition that the prior* π *is a positive and Lipschitz function over* Θ*, that*

$\theta \mapsto |I(\theta)|$ (determinant of $I(\theta)$) and $\theta \mapsto x'I(\theta)x$ (for all $x \in \mathbb{R}^q$, x' being the transpose of x) are Lipschitz functions over Θ, and that the constant which arises in the Lipschitz condition for $\theta \mapsto x'I(\theta)x$ and which is a function of x, is also a Lipschitz function over $\mathbb{R}^q$. Then, when $n \to \infty$, the posterior density $p(\theta \mid \hat{\theta}_{ML})$ conditional on the MLE is asymptotically equivalent to the density of the normal distribution with mean vector $\hat{\theta}_{ML}$ and covariance matrix $n^{-1}I(\hat{\theta}_{ML})^{-1}$ over a subset B_n of Θ whose measure with respect to this normal density is asymptotically one in probability:

$$p(\theta \mid \hat{\theta}_{ML}) \underset{n\to\infty}{\sim} \phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\theta), \qquad \forall \theta \in B_n$$
$$\lim_{n\to\infty} \int_{B_n} \phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\theta) d\theta \underset{\mathbb{P}}{=} 1.$$

Thus, over the subset B_n which asymptotically contains all the mass of the normal density $\phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\cdot)$, the posterior conditional on the MLE is asymptotically equivalent to this normal distribution.

From a frequentist point of view, the BvM theorem (which concerns the full sample posterior $p(\theta \mid \mathcal{D})$) is a justification of the Bayesian approach for parameter estimation since the Bayesian confidence sets asymptotically coincide with the frequentist limit confidence sets (Freedman, 1999). Lemma 1 shows a similar result for the posterior conditional on the MLE $p(\theta \mid \hat{\theta}_{ML})$. Thus, Lemma 1 can also be viewed as a justification of the use of the posterior conditional on asymptotically normal MLE for parameter estimation. Note that results similar to the one provided by Lemma 1 have already been obtained for the estimation of an univariate location parameter; see Doksum and Lo (1990) and references therein.

Regarding assumptions in Lemma 1, the asymptotic normality of the MLE (Equation (1)) requires classical but strong assumptions (even for the simple i.i.d. case). However, the asymptotic normality of the MLE has been

obtained for numerous modeling and sampling settings, even in non-i.i.d. cases. Lemma 1 is also based on a series of Lipschitz assumptions concerning the prior π and the Fisher information matrix $I(\theta)$. These assumptions are satisfied for classical distributions (e.g. when π is uniform on a bounded domain and when $(D_1, \ldots, D_n)$ are independent normal variables with mean μ and variance 1).

Proof. Under the assumptions of the theorem of the MLE asymptotic normality (i.e. under Equation (1)), $p(\hat{\theta}_{ML} \mid \theta) = \eta_n(\theta) + \varepsilon_n(\theta)$, where $\eta_n(\theta) = \phi_{\theta, n^{-1}I(\theta)^{-1}}(\hat{\theta}_{ML})$ and $\varepsilon_n(\theta) \underset{n\to\infty}{=} o(\eta_n(\theta))$. Therefore,

$$p(\theta \mid \hat{\theta}_{ML}) \underset{n\to\infty}{=} \frac{\{\eta_n(\theta) + \varepsilon_n(\theta)\}\pi(\theta)}{\int_\Theta \eta_n(\alpha)\pi(\alpha)d\alpha + \int_\Theta \varepsilon_n(\alpha)\pi(\alpha)d\alpha} = \frac{\eta_n(\theta)\pi(\theta)(1+o(1))}{\int_\Theta \eta_n\pi + \int_\Theta \varepsilon_n\pi}.$$

The densities π and η_n being positive over Θ and the density η_n converging to the Dirac distribution at the true parameter vector θ_0 when $n \to \infty$, $\int_\Theta \eta_n\pi$ is positive and its limit, namely $\pi(\theta_0)$, is also positive. Besides, $\theta \mapsto p(\hat{\theta}_{ML} \mid \theta)$ being asymptotically equivalent to η_n, it also converges to the Dirac distribution at the true parameter vector θ_0 when $n \to \infty$, and $\int_\Theta p(\hat{\theta}_{ML} \mid \theta)\pi(\theta)d\theta \to \pi(\theta_0) > 0$. Therefore, $\int_\Theta \varepsilon_n\pi = \int_\Theta p(\hat{\theta}_{ML} \mid \theta)\pi(\theta)d\theta - \int_\Theta \eta_n\pi$ converges to 0, $\int_\Theta \varepsilon_n\pi = o(\int_\Theta \eta_n\pi)$ and

$$p(\theta \mid \hat{\theta}_{ML}) \underset{n\to\infty}{=} \frac{\pi(\theta)}{\pi(\theta_0)}\eta_n(\theta)(1+o(1)). \tag{2}$$

Let B_n be the ball of center θ_0 and radius r_n such that $r_n \to 0$ and $r_n\sqrt{n} \to \infty$ (r_n converges to zero at a lower rate than $1/\sqrt{n}$). Since π is Lipshitzian (i.e. $\exists A_1 < \infty, \forall \theta_1, \theta_2 \in \Theta, |\pi(\theta_1) - \pi(\theta_2)| \leq A_1||\theta_1 - \theta_2||$),

$$\pi(\theta) \underset{n\to\infty}{=} \pi(\theta_0)(1+o(1)), \qquad \forall \theta \in B_n. \tag{3}$$

Let us now derive an equivalent function for η_n which can be written:

$$\eta_n(\theta) = \phi_{\theta, n^{-1}I(\theta)^{-1}}(\hat{\theta}_{ML}) = \frac{\sqrt{n}|I(\theta)|^{1/2}}{(2\pi)^{d/2}} \exp\left(-\frac{n}{2}(\hat{\theta}_{ML} - \theta)'I(\theta)(\hat{\theta}_{ML} - \theta)\right).$$

Using the Lipschitz condition on $\theta \mapsto I(\theta)$ (i.e. $\exists A_2 < \infty, \forall \theta_1, \theta_2 \in \Theta, |$ $|I(\theta_1)|-|I(\theta_2)| \leq A_2||\theta_1-\theta_2||$), one can state that for all $\theta \in B_n$, $|I(\theta)|^{1/2} \underset{n\to\infty}{=}$ $|I(\hat{\theta}_{ML})|^{1/2}(1+o(1))$. Besides, the two Lipschitz conditions concerning $\theta \mapsto x'I(\theta)x$ can be written as follows:

$$\forall x \in \mathbb{R}^q, \exists A_3(x) < \infty, \forall \theta_1, \theta_2 \in \Theta, |x'I(\theta_1)x - x'I(\theta_2)x| \leq A_3(x)||\theta_1 - \theta_2||$$

$$\exists A_4 < \infty, \forall x_1, x_2 \in \mathbb{R}^q, |A_3(x_1) - A_3(x_2)| \leq A_4||x_1 - x_2||.$$

These conditions imply that over B_n, $|(\hat{\theta}_{ML} - \theta)'I(\theta)(\hat{\theta}_{ML} - \theta) - (\hat{\theta}_{ML} -$ $\theta)'I(\hat{\theta}_{ML})(\hat{\theta}_{ML} - \theta)|$ is bounded from above by $2r_n(A_3(0) + 2r_nA_4)$. Therefore, $(\hat{\theta}_{ML} - \theta)'I(\theta)(\hat{\theta}_{ML} - \theta) \underset{n\to\infty}{=} (\hat{\theta}_{ML} - \theta)'I(\hat{\theta}_{ML})(\hat{\theta}_{ML} - \theta)(1+o(1))$and

$$\eta_n(\theta) \underset{n\to\infty}{=} \phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\theta)(1+o(1)), \qquad \forall \theta \in B_n. \tag{4}$$

Using Equations (2-4), we obtain the first equation of Lemma 1:

$$p(\theta \mid \hat{\theta}_{ML}) \underset{n\to\infty}{=} \phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\theta)(1+o(1)), \qquad \forall \theta \in B_n.$$

The 2nd equation of Lemma 1 is shown below. Let $p \in (0,1)$, and consider $R_{n,p}$ the region consisting of the vectors $\theta \in \Theta$ satisfying $n(\hat{\theta}_{ML} -$ $\theta)'I(\hat{\theta}_{ML})(\hat{\theta}_{ML} - \theta) \leq \chi^2_d(p)$, where $\chi^2_d(p)$ is the quantile of order p of the chi-square distribution of order d (i.e. the dimension of Θ). Using the link between the normal and chi-square distributions, $R_{n,p}$ satisfies: $\int_{R_{n,p}} \phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\theta)d\theta = p$. Moreover, from a property of the Rayleigh quotient, $(\hat{\theta}_{ML} - \theta)'I(\hat{\theta}_{ML})(\hat{\theta}_{ML} - \theta) \geq \lambda_{min}\{I(\hat{\theta}_{ML})\}||\hat{\theta}_{ML} - \theta||$, where $\lambda_{min}\{I(\hat{\theta}_{ML})\}$ is the minimum eigenvalue of $I(\hat{\theta}_{ML})$. Therefore,

$$\forall \theta \in R_{n,p}, \quad ||\hat{\theta}_{ML} - \theta|| \leq \lambda_{min}\{I(\hat{\theta}_{ML})\}^{-1}\sqrt{\chi^2_d(p)/n}. \tag{5}$$

From the MLE asymptotic normality,

$$\lim_{n\to\infty} \mathbb{P}\left((\hat{\theta}_{ML} - \theta_0)'I(\theta_0)(\hat{\theta}_{ML} - \theta_0) \leq \sqrt{\chi^2_d(p)/n}\right) = p.$$

Moreover, $(\hat{\theta}_{ML} - \theta_0)' I(\theta_0)(\hat{\theta}_{ML} - \theta_0) \geq \lambda_{min}\{I(\theta_0)\}||\hat{\theta}_{ML} - \theta_0||$, where $\lambda_{min}\{I(\theta_0)\}$ is the minimum eigenvalue of $I(\theta_0)$. Therefore,

$$\lim_{n\to\infty} \mathbb{P}\left(||\hat{\theta}_{ML} - \theta_0|| \leq \lambda_{min}\{I(\theta_0)\}^{-1}\sqrt{\chi^2_d(p)/n}\right) \geq p. \tag{6}$$

Since $||\theta - \theta_0|| \leq ||\hat{\theta}_{ML} - \theta|| + ||\hat{\theta}_{ML} - \theta_0||$, one obtains using Eq. (5) and (6),

$$\lim_{n\to\infty} \mathbb{P}\left(\forall \theta \in R_{n,p}, ||\theta - \theta_0|| \leq \left[\lambda_{min}\{I(\hat{\theta}_{ML})\}^{-1} + \lambda_{min}\{I(\theta_0)\}^{-1}\right]\sqrt{\chi^2_d(p)/n}\right) \geq p.$$

Since r_n goes to zero more slowly than $1/\sqrt{n}$, the previous inequality yields:

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left(\forall \theta \in R_{n,p}, ||\theta - \theta_0|| \leq r_n\right) &\geq p \\
\lim_{n\to\infty} \mathbb{P}\left(R_{n,p} \subset B_n\right) &\geq p \\
\lim_{n\to\infty} \mathbb{P}\left(\int_{B_n} \phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\theta)d\theta \geq \int_{R_{n,p}} \phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\theta)d\theta\right) &\geq p \\
\lim_{n\to\infty} \mathbb{P}\left(\int_{B_n} \phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\theta)d\theta \geq p\right) &\geq p.
\end{aligned}$$

The last inequality obtained for any $p \in (0,1)$ implies that $\int_{B_n} \phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\theta)d\theta$ converges to one in probability when $n \to \infty$, i.e.

$$\lim_{n\to\infty} \int_{B_n} \phi_{\hat{\theta}_{ML}, n^{-1}I(\hat{\theta}_{ML})^{-1}}(\theta)d\theta \underset{\mathbb{P}}{=} 1.$$

□

4. Posterior conditional on an MPLE

Here, we propose a lemma analogous to Lemma 1 but concerning an MPLE (maximum pseudo-likelihood estimate) instead of the MLE.

Lemma 2 (Asymptotic normality of the posterior conditional on an MPLE)**.** *Consider the modeling setting provided in Section 2. Assume that, given the vector θ under which the data $\mathcal{D}$ were generated, the p.d.f. of the MPLE*

$\hat{\theta}_{MPL}$ is equivalent to the normal density with mean vector θ and covariance matrix $g(n)^{-1}J(\theta)^{-1}$:

$$p(\hat{\theta}_{MPL} \mid \theta) \underset{n\to\infty}{\sim} \phi_{\theta, g(n)^{-1}J(\theta)^{-1}}(\hat{\theta}_{MPL}),$$

*where g is a positive increasing function such that $g(n) \to \infty$ and $J(\theta)$ is a positive-definite matrix. Assume in addition that the prior π is a positive and Lipschitz function over Θ, that $\theta \mapsto |J(\theta)|$ and $\theta \mapsto x'J(\theta)x$ (for all $x \in \mathbb{R}^q$, x' being the transpose of x) are Lipschitz functions over Θ, and that the constant which arises in the Lipschitz condition for $\theta \mapsto x'J(\theta)x$ and which is a function of x, is also a Lipschitz function over $\mathbb{R}^q$. Then, when $n \to \infty$, the posterior density $p(\theta \mid \hat{\theta}_{MPL})$ conditional on the MPLE is asymptotically equivalent to the density of the normal distribution with mean vector $\hat{\theta}_{MPL}$ and covariance matrix $g(n)^{-1}J(\hat{\theta}_{MPL})^{-1}$ over a subset B_n of Θ whose measure with respect to this normal density is asymptotically*one:

$$p(\theta \mid \hat{\theta}_{MPL}) \underset{n\to\infty}{\sim} \phi_{\hat{\theta}_{MPL}, g(n)^{-1}J(\hat{\theta}_{MPL})^{-1}}(\theta), \qquad \forall \theta \in B_n$$
$$\lim_{n\to\infty} \int_{B_n} \phi_{\hat{\theta}_{MPL}, g(n)^{-1}J(\hat{\theta}_{MPL})^{-1}}(\theta) d\theta = 1.$$

Lemma 2 justifies the use of the posterior conditional on the MPLE for parameter estimation because the Bayesian confidence sets that are provided by this posterior asymptotically coincide with the frequentist limit confidence sets obtained by maximizing the pseudo-likelihood.

The asymptotic normality of the MPLE required in Lemma 2 has been obtained for various models, especially random Markov fields and spatial point processes; see Gaetan and Guyon (2008, chap. 5), Gourieroux et al. (1983), Møller and Waagepetersen (2004, chap. 9) and references therein. It has to be noted that information is lost when MPLE are used rather than MLE and, consequently, that estimation accuracy is decreased (e.g. this

has been shown for simple Markovian models using asymptotic estimation variances (Gaetan and Guyon, 2008, chap. 5)).

Proof. Follow the proof of Lemma 1 by assuming that the radius r_n of the ball B_n satisfies $r_n\sqrt{g(n)} \to \infty$ instead of $r_n\sqrt{n} \to \infty$.

□

5. Approximate posterior conditional on an MPLE

Here, we derive implications of Lemma 2 in the framework of approximate Bayesian computation (ABC) when (some of) the summary statistics are MPLE (see Theorem 1 and Corollary 1). We consider the (simple) ABC–rejection algorithm based on independent simulations, on a set of summary statistics and on a tolerance threshold (Pritchard et al., 1999):

ABC–rejection. Perform the next 3 steps for i in $\{1, \ldots, I\}$, independently:

- Generate θ_i from π and simulate $\mathcal{D}_i$ from $\mathcal{M}_{\theta_i}$;
- Compute summary statistics $S_i = s(\mathcal{D}_i)$, where s is a function from $\mathbb{D}$ to the space $\mathbb{S}$ of statistics;
- Accept θ_i if $d(S_i, S) \leq \epsilon$, where d is a distance over $\mathbb{S}$ and ϵ is a tolerance threshold for the distance between the observed statistics $S = s(\mathcal{D})$ and the simulated ones S_i.

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

$$p_\epsilon(\theta \mid S) = \frac{\left(\int_{\mathcal{B}_d(S,\epsilon)} h(s \mid \theta) ds\right) \pi(\theta)}{\int_\Theta \left(\int_{\mathcal{B}_d(S,\epsilon)} h(s \mid \alpha) ds\right) \pi(\alpha) d\alpha},$$

where the ball $\mathcal{B}_d(S, \epsilon)$ in the d-dimensional space $\mathbb{S}$ is the set of points from which the distance to S is less than ϵ, and $h(S \mid \theta)$ is the conditional

probability distribution function of S given θ. When ϵ tends to zero and d is appropriate, $p_\epsilon(\theta \mid S)$ is a good approximation of $p(\theta \mid S)$ under regularity assumptions (see Blum (2010) and Soubeyrand et al. (2013, Appendix A)): $p_\epsilon(\theta \mid S)$ and $p(\theta \mid S)$ are asymptotically equivalent. However, if S is not sufficient, then $p(\theta \mid S) \neq p(\theta \mid \mathcal{D})$ and information is lost by using S instead of $\mathcal{D}$.

Theorem 1 (Asymptotic normality of the ABC–posterior conditional on an MPLE)**.** *Consider the ABC–rejection algorithm that samples in the posterior $p_\epsilon(\theta \mid \hat{\theta}_{MPL})$ of θ conditional on the vector of summary statistics $S = \hat{\theta}_{MPL}$. Assume that when $\epsilon \to 0$, $p_\epsilon(\theta \mid \hat{\theta}_{MPL})$ converges pointwise to $p(\theta \mid \hat{\theta}_{MPL})$. Then, under assumptions of Lemma 2, when $n \to \infty$ and $\epsilon \to 0$, the posterior $p_\epsilon(\theta \mid \hat{\theta}_{MPL})$ is asymptotically equivalent to the density of the normal distribution with mean vector $\hat{\theta}_{MPL}$ and covariance matrix $g(n)^{-1} J(\hat{\theta}_{MPL})^{-1}$ over a subset B_n of Θ whose measure with respect to this normal density goes to one in probability and that does not depend on ϵ:*

$$p_\epsilon(\theta \mid \hat{\theta}_{MPL}) \underset{n\to\infty,\epsilon\to 0}{\sim} \phi_{\hat{\theta}_{MPL}, g(n)^{-1} J(\hat{\theta}_{MPL})^{-1}}(\theta), \qquad \forall \theta \in B_n$$

$$\lim_{n\to\infty} \int_{B_n} \phi_{\hat{\theta}_{MPL}, g(n)^{-1} J(\hat{\theta}_{MPL})^{-1}}(\theta) d\theta \underset{\mathbb{P}}{=} 1.$$

As explained in the introduction, this result leads us to conjecture that, for some stochastic implicit models, (i) the ABC–posterior distribution conditional on an MPLE is asymptotically normal and centered around the MPLE, and (ii) resulting point estimates and confidence sets converge to their frequentist analogues. Corollary 1 in Appendix A provides an analogous result when the MPLE is used in conjunction with supplementary statistics.

Proof. This result is directly obtained from Lemma 2 by simply noting that

the subset B_n does not depend on ϵ. Indeed, following the proof of Lemma 1, pointwise in θ we have:

$$p_\epsilon(\theta \mid \hat{\theta}_{MPL}) \underset{\epsilon \to 0}{=} p(\theta \mid \hat{\theta}_{MPL})(1 + o_\epsilon(1)) \\ \underset{n \to \infty, \epsilon \to 0}{=} \frac{\pi(\theta)}{\pi(\theta_0)} \eta_n(\theta)(1 + o_n(1))(1 + o_\epsilon(1)).$$

Then, B_n is used to provide an equivalent of $\pi(\theta)\pi(\theta_0)^{-1}\eta_n(\theta)$. This term does not depend on ϵ and, consequently, B_n has not to be dependent on ϵ.

□

6. Example

The simplified example presented here illustrates the application of ABC conditional on an MPLE and a supplementary statistic. The model $\mathcal{M}_\theta$ under consideration is the following bivariate normal distribution:

$$\mathcal{N}\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right),$$

parameterized by the mean μ and the correlation ρ; we set $\theta = (\mu, \rho)$. Observed data $\mathcal{D} = \{(D_k^{(1)}, D_k^{(2)}) : k = 1, \ldots, n\}$ are $n = 100$ vectors drawn under this normal distribution with $\mu = 0$ and $\rho = 0.5$. We use a uniform prior distribution π over the rectangular domain $(-3, 3) \times (-1, 1)$. The maximum likelihood estimates of μ and ρ are the empirical mean of $(D_k^{(1)} + D_k^{(2)})/2$ and the empirical correlation of $(D_k^{(1)}, D_k^{(2)})$, $k = 1, \ldots, n$. Here, we applied ABC with the two following statistics:

$$S = s(\mathcal{D}) = \begin{pmatrix} \hat{\mu}_{MPL} \\ S_0 \end{pmatrix} = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} D_k^{(1)} \\ \mathbf{1}\{\text{sign}(D_k^{(1)}) = \text{sign}(D_k^{(2)})\} \end{pmatrix},$$

where $\hat{\mu}_{MPL}$ is an MPLE of μ that uses only partial information contained in the sample (i.e. only the first component of sampled vectors), and S_0 is a

supplementary statistic that gives the mean number of vectors in the sample whose components $D_k^{(1)}$ and $D_k^{(2)}$ have the same sign ($\mathbf{1}\{\cdot\}$ is the indicator function).

To assess the convergence of ABC when ϵ tends to zero, we applied ABC with varying ϵ, with $I = 10^5$ simulations, and with the distance $d(S_i, S) = (\hat{\mu}_{MPL,i} - \hat{\mu}_{MPL})^2 + (S_{0,i} - S_0)^2$, where $S_i = (\hat{\mu}_{MPL,i}, S_{0,i})$ is the vector of statistics computed for the simulation i. As usual in ABC–rejection, instead of fixing ϵ, we fixed the sample size τ of the posterior sample (i.e. the number of accepted parameter vectors); note that ϵ decreases when τ decreases. The sample size τ was fixed at values ranging from 10 to 5000. For each value of τ, we computed the local posterior probability (LPP) around the true parameter vector $\theta = (0, 0.5)$ as the proportion of accepted parameter vectors in the small rectangle $[-0.015, 0.015] \times [0.495, 0.505]$ whose center is $\theta = (0, 0.5)$ and whose sides are 200 times smaller than the sides of the parameter space $(-3, 3) \times (-1, 1)$. We expect that this LPP increases with the efficiency of the inference procedure. The LPP was computed for 50000 datasets and Figure 1 shows its mean and standard deviation when τ varies. The mean LPP around the true parameters increases when the sample size τ (and ϵ) tends to zero; meanwhile, the dispersion of the LPP increases. This is the signature of the classical bias–variance trade-off.

To automatically select the sample size τ, we applied the procedure proposed by Soubeyrand et al. (2013) where the distance between summary statistics is also optimized. In this procedure, the distance is weighted: $d(S_i, S; w_1, w_2) = w_1(\hat{\mu}_{MPL,i} - \hat{\mu}_{MPL})^2 + w_2(S_{0,i} - S_0)^2$, and the triplet (τ, w_1, w_2) is optimized under constraints using an integrated mean square error criterion. For one of the 50000 datasets simulated above, Figure 2 shows ABC–posterior samples obtained when d is not weighted and τ is

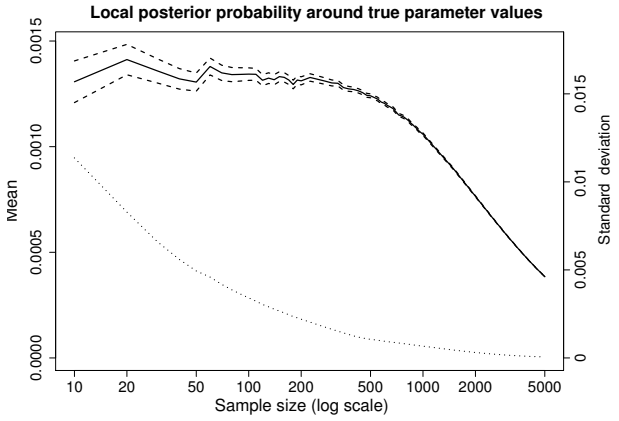

Figure 1: Mean (solid line), pointwise 95%-confidence envelopes (dashed lines) and standard deviation (dotted line) of the local posterior probability around the true parameter vector $\theta = (0, 0.5)$ as a function of sample size τ.

fixed at 5000, 1000 and 200, and when d is weighted and (τ, w_1, w_2) is optimized. The grey contour line shows the smallest 95%-posterior area obtained with the classical Bayesian computation. The first three panels illustrate the bias–variance trade-off when τ tends to zero. The fourth panel illustrates the difference between the classical Bayesian inference conditional on all data and the ABC inference conditional on partial information and with optimized τ (here $\tau = 585$). The relevancy of the optimized sample size $\tau = 585$ can be seen by projecting this value on Figure 1: (i) the expected LPP around the true parameters is comparable to expected LPP obtained with lower τ, and (ii) the standard deviation of the LPP is strongly decreased compared with standard deviations obtained with lower τ.

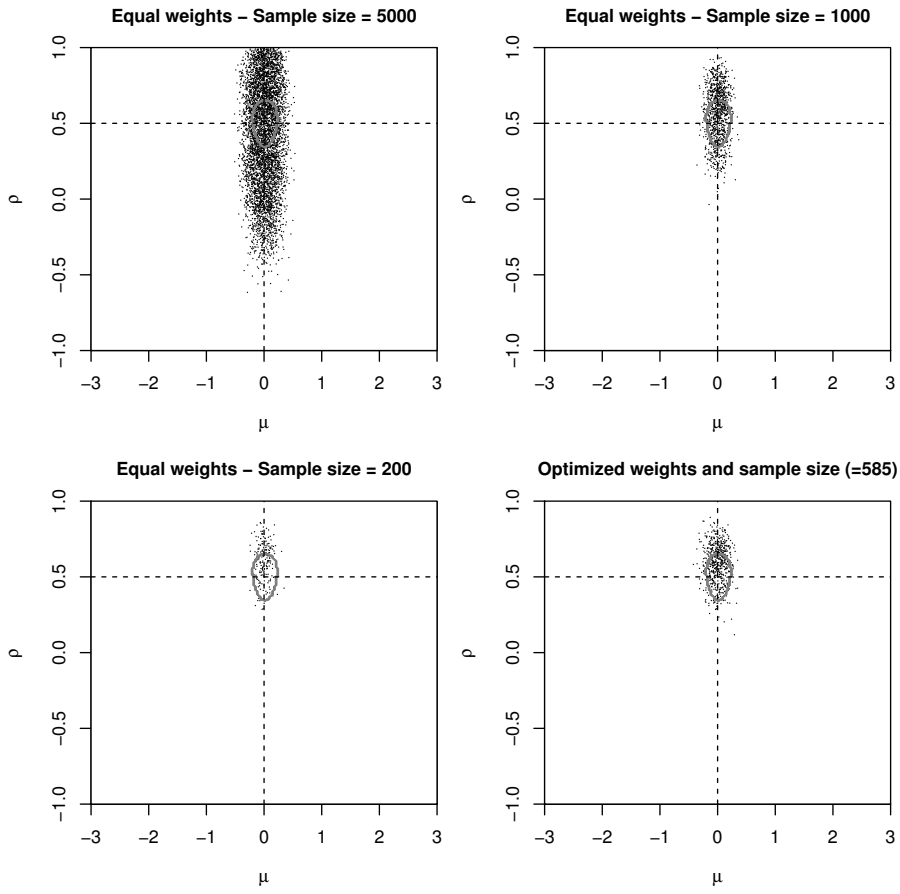

Figure 2: ABC–posterior samples (dots) obtained when d is not weighted and τ is fixed at 5000 (top left), 1000 (top right) and 200 (bottom left), and when d is weighted and (τ, w_1, w_2) is optimized (bottom right). Dashed lines are intersecting at the true value $(0, 0.5)$ of the parameter vector $\theta = (\mu, \rho)$. The grey contour line gives the smallest 95%-posterior area obtained with the classical Bayesian computation.

7. Discussion

We provided a frequentist justification for the use of posterior distributions conditional on MPLE, both in the classical Bayesian framework and in the ABC framework. The asymptotic results presented above were obtained for a large but limited class of models satisfying regularity assumptions. In real-life studies where ABC is applied, these assumptions cannot be checked and, consequently, our asymptotic results may not hold. Therefore, it is crucial (i) to combine MPLE and supplementary summary statistics to tend to a set of sufficient summary statistics (Joyce and Marjoram, 2008), and (ii) to apply a method for selecting, weighting or transforming the summary statistics to avoid to take into account non-relevant statistics. In Section 6, we used the weighted distance between summary statistics proposed by Soubeyrand et al. (2013) where the weights are optimized with respect to an integrated mean square error. Other approaches could be applied. For example, Barnes et al. (2012), Joyce and Marjoram (2008) and Nunes and Balding (2010) propose a dimension reduction (that can be viewed as a binary weighting), and Wegmann et al. (2009) proposes a PLS transformation followed by a binary weighting of the PLS axes (see also Blum et al., 2013; Fearnhead and Prangle, 2012; Jung and Marjoram, 2011).

The strong implication of our approach is that an analytic work has to be made: the dependence structure of the model has to be simplified to write a tractable pseudo-likelihood and, eventually, to find an analytic expression for the maximizer. This additional work is however expected to yield relevant summary statistics directly informing (a subset of) the parameters.

In this article, we did not precisely define what is a pseudo-likelihood. This deliberate choice is justified in the sense that the MPLE could be

replaced, in Theorem 1, by any other estimates with similar normal weak convergence. Thus, our approach could be generalized by replacing the pseudo-likelihood by a quasi-likelihood or a contrast. Another extension of Theorem 1 could be derived by conditioning ABC on the score function of the pseudo-likelihood instead of the MPLE as proposed by Gleim and Pigorsch (2013), Ruli et al. (2013) and Mengersen et al. (2013).

Acknowledgements. We thank Denis Allard, Rachid Senoussi and the reviewers for their suggestions. This work was supported by the ANR grant EMILE.

Appendix A. Combining an MPLE and supplementary statistics

Corollary 1. *Consider the posterior $p_\epsilon(\theta \mid \hat{\theta}_{MPL}, S_0)$ of θ conditional on the vector of summary statistics $S = (\hat{\theta}_{MPL}, S_0)$ in which the ABC–rejection algorithm samples. Assume that when $\epsilon \to 0$, $p_\epsilon(\theta \mid \hat{\theta}_{MPL})$ converges pointwise to $p(\theta \mid \hat{\theta}_{MPL})$. Assume that there exists a vector T_0 independent of $\hat{\theta}_{MPL}$ such that $(\hat{\theta}_{MPL}, T_0)$ brings the same information on θ than $S = (\hat{\theta}_{MPL}, S_0)$. Then,*

$$p_\epsilon(\theta \mid \hat{\theta}_{MPL}, S_0) \underset{\epsilon \to 0}{=} \check{p}(\theta \mid T_0)(1 + o_\epsilon(1)),$$

where $\check{p}(\theta \mid T_0)$ is the posterior density of θ conditional on the subset of statistics T_0 given the modified prior $p(\theta \mid \hat{\theta}_{MPL}) = p(\hat{\theta}_{MPL} \mid \theta)\pi(\theta)/p(\hat{\theta}_{MPL})$.

If in addition assumptions of Lemma 2 hold, the modified prior $p(\theta \mid \hat{\theta}_{MPL})$ is asymptotically equivalent to the density of the normal distribution with mean vector $\hat{\theta}_{MPL}$ and covariance matrix equal to $g(n)^{-1} J(\hat{\theta}_{MPL})^{-1}$ over a subset B_n of Θ whose measure with respect to this normal density goes to one in probability and that does not depend on ϵ.

Under the first set of assumptions of Corollary 1, we can also write that $p_\epsilon(\theta \mid \hat{\theta}_{MPL}, S_0) \underset{\epsilon \to 0}{=} \check{p}(\theta \mid \hat{\theta}_{MPL})(1 + o_\epsilon(1))$, where $\check{p}(\theta \mid \hat{\theta}_{MPL})$ is the posterior density of θ conditional on the MPLE with the modified prior $p(\theta \mid T_0) = p(T_0 \mid \theta)\pi(\theta)/p(T_0)$. Moreover, the existence of T_0 is a strong assumption that reduces the applicability of the corollary (a trivial but certainly rare example where T_0 exists is obtained when one assumes that $(\hat{\theta}_{MPL}, S_0)$ is a normal random vector; Indeed, in this case T_0 can be defined as $T_0 = S_0 - E(S_0 \mid \hat{\theta}_{MPL})$). Nevertheless, if $(\hat{\theta}_{MPL}, S_0)$ is asymptotically normal, we conjecture that a result similar to Corollary 1 can be obtained.

Proof. The first part of the corollary is straightforward:

$$
\begin{aligned}
p_\epsilon(\theta \mid \hat{\theta}_{MPL}, S_0) &\underset{\epsilon \to 0}{=} p(\theta \mid \hat{\theta}_{MPL}, S_0)(1 + o_\epsilon(1)) = p(\theta \mid \hat{\theta}_{MPL}, T_0)(1 + o_\epsilon(1)) \\
&= \frac{p(\hat{\theta}_{MPL}, T_0 \mid \theta)\pi(\theta)}{p(\hat{\theta}_{MPL}, T_0)}(1 + o_\epsilon(1)) = \frac{p(T_0 \mid \hat{\theta}_{MPL}, \theta)p(\hat{\theta}_{MPL} \mid \theta)\pi(\theta)}{p(T_0)p(\hat{\theta}_{MPL})}(1 + o_\epsilon(1)) \\
&= \frac{p(T_0 \mid \theta)p(\hat{\theta}_{MPL} \mid \theta)\pi(\theta)}{p(T_0)p(\hat{\theta}_{MPL})}(1 + o_\epsilon(1)) = \frac{p(T_0 \mid \theta)p(\theta \mid \hat{\theta}_{MPL})}{p(T_0)}(1 + o_\epsilon(1)).
\end{aligned}
$$

The term $p(\theta \mid \hat{\theta}_{MPL})$ can be viewed as a prior that is based on information contained in $\hat{\theta}_{MPL}$ and that is independent from data T_0 used for the final inference. The second part of the corollary corresponds to Theorem 1.

□

Barnes, C. P., Filippi, S., Stumpf, M. P. H., Thorne, T., 2012. Considerate approaches to constructing summary statistics for abc model selection. Statistics and Computing 6, 1181–1197.

Bickel, P. J., Kleijn, B. J. K., 2012. The semiparametric Bernstein–von Mises theorem. The Annals of Statistics 40, 206–237.

Blum, M. G. B., 2010. Approximate Bayesian computation: A nonparametric perspective. J. Amer. Statist. Assoc. 105, 1178–1187.

Blum, M. G. B., Nunes, M. A., Prangle, D., Sisson, S. A., 2013. A comparative review of dimension reduction methods in approximate Bayesian computation. Statistical Science 28, 189–208.

Bochkina, N. A., Green, P. J., 2014. The bernstein–von Mises theorem and nonregular models. The Annals of Statistics 42, 1850–1878.

Bontemps, D., 2011. Bernstein–von Mises theorems for Gaussian regression with increasing number of regressors. The Annals of Statistics 39, 2557–2584.

Borwanker, J., Kallianpur, G., Prakasa Rao, B. L. S., 1971. The Bernstein-von Mises theorem for Markov processes. The Annals of Mathematical Statistics, 1241–1253.

Castillo, I., 2012a. Semiparametric bernstein–von mises theorem and bias, illustrated with gaussian process priors. Sankhya A 74, 194–221.

Castillo, I., 2012b. A semiparametric Bernstein–von Mises theorem for Gaussian process priors. Probability Theory and Related Fields 152, 53–99.

Castillo, I., Nickl, R., 2013. Nonparametric Bernstein–von Mises theorems in Gaussian white noise. The Annals of Statistics 41, 1999–2028.

Doksum, K. A., Lo, A. Y., 1990. Consistent and robust Bayes procedures for location based on partial information. Ann. Stat. 18, 443–453.

Drovandi, C. C., Pettitt, A. N., Faddy, M. J., 2011. Approximate Bayesian

computation using indirect inference. Journal of the Royal Statistical Society C 60, 317–337.

Fearnhead, P., Prangle, D., 2012. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. J. Roy. Stat. Soc. B 74, 419–474.

Freedman, D., 1999. On the Bernstein-Von Mises theorem with infinite-dimensional parameters. Ann. Stat. 27, 1119–1140.

Gaetan, C., Guyon, X., 2008. Modélisation et statistique spatiale. Springer-Verlag, Berlin.

Gleim, A., Pigorsch, C., 2013. Approximate Bayesian computation with indirect summary statistics. preprint.

Gourieroux, C., Monfort, A., Trognon, A., 1983. Pseudo maximum likelihood methods: Theory. Econometrica 52, 681–700.

Joyce, P., Marjoram, P., 2008. Approximately sufficient statistics and Bayesian computation. Statistical Applications in Genetics and Molecular Biology 7, 1–16.

Jung, H., Marjoram, P., 2011. Choice of summary statistic weights in approximate Bayesian computation. Statistical Applications in Genetics and Molecular Biology 10, 1–23.

Kleijn, B. J. K., van der Vaart, A. W., 2012. The bernstein-von-Mises theorem under misspecification. Electronic Journal of Statistics 6, 354–381.

Lehmann, E. L., Casella, G., 1998. Theory of point estimation, 2nd edition. Vol. 31. Springer.

Lindley, D. V., 1965. Introduction to probability and statistics from Bayesian viewpoint. Part 2: inference. Cambridge University Press.

Marin, J. M., Pudlo, P., Robert, C. P., Ryder, R. J., 2012. Approximate Bayesian computational methods. Stat. Comput. 22, 1167–1180.

Mengersen, K. L., Pudlo, P., Robert, C. P., 2013. Bayesian computation via empirical likelihood. PNAS 110, 1321–1326.

Møller, J., Waagepetersen, R. P., 2004. Statistical Inference and Simulation for Spatial Point Processes. Chapman & Hall/CRC, Boca Raton.

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

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., Feldman, M. W., 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Molecular Biology and Evolution 16, 1791–1798.

Rivoirard, V., Rousseau, J., 2012. Bernstein–von Mises theorem for linear functionals of the density. The Annals of Statistics 40, 1489–1523.

Ruli, E., Sartori, N., Ventura, L., 2013. Approximate Bayesian computation with composite score functions. arXiv:1311.7286.

Soubeyrand, S., Carpentier, F., Desassis, N., Chadoeuf, J., 2009. Inference with a contrast-based posterior distribution and application in spatial statistics. Statistical Methodology 6, 466–477.

Soubeyrand, S., Carpentier, F., Guiton, F., Klein, E. K., 2013. Approximate

Bayesian computation with functional statistics. Stat. App. Genet. Mol. 12, 17–37.

Tamaki, K., 2008. The Bernstein-von Mises theorem for stationary processes. J. Japan Statist. Soc 38, 311–323.

Walker, A. M., 1969. On the asymptotic behaviour of posterior distributions. J. Roy. Stat. Soc. B 31, 80–88.

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