Approximate Bayesian Computation
with composite score functions

Erlis Ruli, Nicola Sartori and Laura Ventura

Department of Statistics, University of Padova, Italy

ruli@stat.unipd.it, sartori@stat.unipd.it, ventura@stat.unipd.it

May 23, 2014

Abstract

Both Approximate Bayesian Computation (ABC) and composite likelihood methods are useful for Bayesian and frequentist inference, respectively, when the likelihood function is intractable. By merging these two methodologies for complex models, we show that composite likelihood score functions can be fruitfully used as automatic informative summary statistics in ABC in order to obtain accurate approximations to the posterior distribution of the parameter of interest. This is formally motivated by the use of the score function of the full likelihood, and extended to general unbiased estimating functions in complex models. In particular, we show that if the composite score is suitably standardized, the simulation scheme is invariant to reparameterizations and automatically adjusts the curvature of the composite likelihood and of the corresponding posterior distribution. Examples illustrate that the proposed ABC procedure can sometimes significantly improve upon usual ABC methods based on ordinary data summaries.

Keywords: Complex models; Composite marginal likelihood; Intractable likelihood; Likelihood-free inference; Pairwise likelihood; Summary statistic; Tangent exponential model; Unbiased
1 Introduction

The summary of the data on a given model offered by the likelihood function is the key ingredient of all likelihood-based inferential methods. However, likelihood-based inference, both frequentist and Bayesian, cannot be performed when the likelihood function is analytically or computationally intractable. This usually occurs in the presence of complex models, such as models with complicated dependence structures or in models with many latent variables.

When the full likelihood is intractable, it is possible to resort to pseudo-likelihood functions, which are intended as surrogates of the full likelihood. An important contribution is given by composite likelihoods, which are based on the composition of suitable lower dimensional densities, such as bivariate marginal (Cox & Reid, 2004), conditional or full conditional densities (Varin et al., 2011). The use of composite likelihoods has been widely advocated in different complex applications of frequentist (see Varin et al., 2011, for a general review, and Larribe & Fearnhead, 2011, for a review in genetics) and of Bayesian inference (Smith & Stephenson, 2009; Pauli et al., 2011; Ribatet et al., 2012).

Even when the computation of the likelihood is impracticable, it is often easy to simulate from the model. Then, an alternative approach to inference may be based on simulations from the model for different parameter values, and on the comparison of simulated datasets with the observed data. The idea is to estimate the likelihood of a given parameter value from the portion of datasets, simulated using that parameter value, that are “similar” to the observed one. This idea was first advocated by Diggle & Gratton (1984).

Approximate Bayesian Computation (ABC) methods combine Diggle & Gratton’s idea with a prior to produce an approximate posterior, which we shall refer to as the ABC posterior (see Beaumont, 2010; Marin et al., 2012). In most applications, the probability of an exact match of the simulated data with the observed data is negligible or zero, hence, the most popular approach is to consider an approximate matching of some summary statistics, evaluated at the observed and simulated data, by means of suitable distances. This method leads to the exact posterior distribution as the distance tends to zero, when the statistics are
sufficient for the parameters of the model. However, in many applications sufficient statistics are not available and the practitioner must resort to a careful selection of data summaries, which could be challenging.

We show that composite likelihoods and ABC ideas can be fruitfully merged in order to obtain accurate approximations to the posterior distribution of the parameter of interest. In particular, we discuss an approach based on composite likelihood score functions for automatically choosing informative summary statistics. This is formally motivated by the use of score function when the full likelihood is available, and is then extended to the use of unbiased estimating functions in complex models. We discuss three examples in which the estimating function is the composite score function. We show empirically that this choice of summary statistic for ABC can sometimes significantly improve upon usual ABC methods based on ordinary data summaries.

The proposed ABC algorithm based on composite score functions (ABC-cs) searches for parameter values of the model of interest that produce simulated data which lead to composite score values – at the observed maximum composite likelihood estimate – close to those based on the original data. This approach has several advantages. First of all, there are as many summary statistics as the number of parameters, and all of them inherit, by construction, useful characteristics of the model. Moreover, the composite score function is generally easy to compute and often it is available analytically. Although composite likelihoods typically do not satisfy the information identity, which leads to wrongly too concentrated posterior distributions \cite{Pauli2011}, if the composite score is suitably standardized, then the simulation scheme is proved to automatically give correctly adjusted posterior approximations and also to be invariant to reparameterizations. Moreover, in the ABC-cs framework and with a small additional computational effort, it is possible to compute the Godambe information, which can be readily used as a precision matrix for more advanced Monte Carlo schemes for ABC (see below).

There have been other attempts to merge composite likelihoods with the ABC framework. For instance, \cite{Erhardt2012}, in the context of spatial extremes, combine composite likelihoods with ABC and show that this approach tends to work better than other existing methods. \cite{Mengersen2013} use the composite score function with the empirical like-
lihood to produce an approximate and weighted posterior sample. However, their approach is not in the framework of typical ABC, as it does not simulate from the full model. Finally, Barthelmé & Chopin (2014, Sec. 7.1) mention the use of composite likelihoods within their ABC approach, based on the Expectation Propagation technique (Minka, 2001), to reduce the computational complexity, although not using the composite score as a summary statistic.

Our approach is similar in spirit with the indirect inference framework (see Heggland & Frigessi, 2004; Gourieroux et al., 1993), as also the ABC-cs method in some sense relies on an auxiliary model likelihood, that is the composite likelihood. As happens in indirect inference, the closer the auxiliary model to the full model the more efficient the parameter estimates will be. However, unlike indirect inference methods, the ABC-cs approach is less computationally demanding since it does not require repeated maximization for each simulated dataset. The indirect inference method within ABC has been discussed by Drovandi et al. (2011).

In Section 2 some background on ABC and composite likelihood methods is given. The proposed ABC algorithm, based on full and composite score functions, is presented in Section 3. Section 4 discusses three examples with various levels of complexity, and a concluding discussion is given in Section 5.

2 Statistical methods

2.1 ABC algorithms

Let $\pi(\theta)$ be a prior distribution for the parameter $\theta \in \Theta \subseteq \mathbb{R}^d$, $L(\theta) = L(\theta; y) = f(y; \theta)$ the likelihood function based on data $y$ and $\pi(\theta|y) \propto \pi(\theta)L(\theta)$ the posterior distribution of $\theta$. Suppose that $L(\theta)$ is unavailable for mathematical or computational reasons.

The primary purpose of ABC algorithms is to approximate the posterior distribution, when usual methods, such as Markov chain Monte Carlo (MCMC), data augmentation, importance sampling or Laplace approximation, cannot be used, but when the data from $f(y; \theta)$ can be easily simulated. The original accept-reject ABC algorithm works by first drawing a candidate parameter value $\theta^*$ from the prior. Then a new dataset $y$ is drawn from the model with the parameter equal to $\theta^*$. Finally, if the simulated data $y$ are equal to
the observed one \( y_{\text{obs}} \), \( \theta^* \) is accepted. With continuous data the equality of \( y \) and \( y_{\text{obs}} \) will happen with probability zero. Hence, in the ABC accept-reject algorithm the exact matching is typically replaced by the condition \( \rho(\eta(y), \eta(y_{\text{obs}})) \leq \epsilon \) (Algorithm 1), where \( \eta(\cdot) \) is a set of suitable summary statistics (e.g. moments, quantiles), \( \rho(\cdot, \cdot) \) is a distance function (e.g. Euclidean distance, absolute norm), and \( \epsilon \) a tolerance threshold.

**Result:** A sample \( (\theta^{(1)}, \ldots, \theta^{(m)}) \) from \( \pi(\theta|\eta(y_{\text{obs}})) \)

```latex
\begin{algorithm}
\textbf{for} i = 1 \rightarrow m \textbf{ do} \\
\quad \textbf{repeat} \\
\quad\quad 1 \hspace{1em} \text{draw } \theta^* \sim \pi(\theta) \hspace{1em} \\
\quad\quad 2 \hspace{1em} \text{draw } y \sim f(y; \theta^*) \hspace{1em} \\
\quad\quad \textbf{until} \hspace{1em} \rho(\eta(y), \eta(y_{\text{obs}})) \leq \epsilon; \hspace{1em} \\
\quad \text{set } \theta^{(i)} = \theta^* \\
\end{algorithm}
```

Algorithm 1: ABC accept-reject sampler.

Algorithm 1 samples from the joint distribution

\[
\pi_\epsilon(\theta, y|\eta(y_{\text{obs}})) = \frac{\pi(\theta) f(y; \theta) \mathbb{1}_{A_{\epsilon,y_{\text{obs}}}}(y)}{\int_{A_{\epsilon,y_{\text{obs}}}} \pi(\theta) f(y; \theta) dy d\theta},
\]

where \( \mathbb{1}_{A_{\epsilon,y_{\text{obs}}}}(y) \) is the indicator function of the set \( A_{\epsilon,y_{\text{obs}}}(y) = \{ y : \rho(\eta(y), \eta(y_{\text{obs}})) \leq \epsilon \} \), and it produces an approximation to the posterior distribution \( \pi(\theta|y_{\text{obs}}) \), given by

\[
\pi_\epsilon(\theta|\eta(y_{\text{obs}})) = \int \pi_\epsilon(\theta, y|\eta(y_{\text{obs}})) dy.
\]

If \( \epsilon \rightarrow 0 \), then \( \pi_\epsilon(\theta|\eta(y_{\text{obs}})) \rightarrow \pi(\theta|\eta(y_{\text{obs}})) \). In addition, if \( \eta(\cdot) \) is sufficient, then \( \pi_\epsilon(\theta|\eta(y_{\text{obs}})) \rightarrow \pi(\theta|y_{\text{obs}}) \) (see, for instance, Marin et al., 2012).

In this respect, ABC suffers from three sources of approximation error: \( \epsilon \), \( \eta(\cdot) \), and the Monte Carlo error. The threshold \( \epsilon \) cannot be fixed to zero, for computational efficiency, and is generally set to the \( \alpha \)th quantile of the distance among the statistics, with \( \alpha \) typically very small (see, for instance, Fearnhead & Prangle, 2012). With non-informative priors, the original accept-reject algorithm may be very inefficient, e.g. the Monte Carlo error may be overwhelming, since simulations from \( \pi(\theta) \) do not account for the data at the
proposal stage, and thus may lead to proposed values located in low posterior probability regions (Marin et al., 2012). Nevertheless, this issue can be effectively addressed by using more advanced Monte Carlo algorithms, such as MCMC methods (Marjoram et al., 2003), importance sampling (Fearnhead & Prangle, 2012), sequential or population Monte Carlo approaches (Sisson et al., 2007, 2009; Beaumont et al., 2009). Hence, in the end, the most crucial point of the ABC algorithm is the choice of $\eta(\cdot)$. Indeed, what ABC can achieve at best is $\pi(\theta|y^{\text{obs}})$, since $\eta(\cdot)$ is rarely sufficient. This loss of information seems to be a necessary price to pay for the access to computable quantities.

### 2.2 Composite likelihoods

In various modern applications likelihood-based methods may encounter computational problems, due to the difficulty, or even impossibility, of evaluating the full likelihood function. In these situations, it is possible to resort to pseudo-likelihoods, called composite likelihoods, which are based on the composition of suitable lower dimensional densities, such as marginal or conditional densities or even a combination of them.

Let $y = (y_1, \ldots, y_n)$ be independent observations from $Y_i \sim f(y_i; \theta)$, where $y_i \in \mathcal{Y} \subseteq \mathbb{R}^q$, and let $\{A_1(y_i), \ldots, A_K(y_i)\}$ be a set of marginal or conditional events on $\mathcal{Y}$, for which the likelihood contribution $L_k(\theta; y_i) \propto f(y \in A_k(y_i); \theta)$ can be computed. The composite log-likelihood is defined as

$$c\ell(\theta; y) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_k \log L_k(\theta; y_i), \quad (2)$$

where $w_k$, $k = 1, \ldots, K$, are non-negative weights. When the events $A_k(y_i)$ are defined in terms of pairs of bivariate marginal densities $f_{hk}(\cdot, \cdot; \theta)$, the associated composite log-likelihood is called pairwise log-likelihood and is given by

$$p\ell(\theta; y) = \sum_{i=1}^{n} \sum_{h,k=1}^{q} \sum_{h \neq k} w_{hk} \log f_{hk}(y_{ih}, y_{ik}; \theta). \quad (3)$$

In some circumstances, it may be useful to consider larger subsets, such as triplets of observations, as in Example 4.
The validity of inference about \( \theta \) using composite likelihoods can be assessed from the standpoint of unbiased estimating functions or the Kullback-Leibler criterion (Lindsay, 1988; Cox & Reid, 2004; Lindsay et al., 2011; Varin et al., 2011). Under rather broad assumptions (see, for instance, Molenberghs & Verbeke, 2005), the maximum composite likelihood estimator (MCLE) \( \hat{\theta} \) is the solution of the composite score equation

\[

c_{\ell\theta}(\theta; y) = \frac{\partial c_{\ell\theta}(\theta; y)}{\partial \theta} = 0.
\]

The composite score \( c_{\ell\theta}(\theta; y) \) is unbiased, i.e. \( E_{\theta}\{c_{\ell\theta}(\theta; Y)\} = 0 \), since it is a linear combination of valid score functions. Moreover, \( \hat{\theta} \) is consistent and approximately normal, with mean \( \theta \) and variance

\[
V(\theta) = H(\theta)^{-1} J(\theta) H(\theta)^{-1},
\]

where \( H(\theta) = E_{\theta}\{-\partial c_{\ell\theta}(\theta; Y) / \partial \theta^T\} \) and \( J(\theta) = \text{var}_{\theta}\{c_{\ell\theta}(\theta; Y)\} \) are the sensitivity and the variability matrices, respectively. The matrix \( G(\theta) = V(\theta)^{-1} \) is known as the Godambe information, and the sandwich form of \( V(\theta) \) is due to the failure of the information identity since, in general, \( H(\theta) \neq J(\theta) \). This failure typically implies that the composite likelihood is wrongly too concentrated.

The asymptotic distribution of the composite likelihood ratio \( cw(\theta) = 2\{c_{\ell\theta}(\hat{\theta}; y) - c_{\ell\theta}(\theta; y)\} \) is a linear combination of independent chi-square random variables, i.e. \( cw(\theta) \sim \sum_{j=1}^{d} \omega_j Z_j^2 \), where \( Z_1, \ldots, Z_d \) are independent standard normal variates and the coefficients \( \omega_1, \ldots, \omega_d \) are the eigenvalues of the matrix \( J(\theta) H(\theta)^{-1} \). In the special case \( d = 1 \), we have \( \omega_1 = J(\theta)/H(\theta) \), so that the adjusted pairwise log-likelihood ratio statistic \( cw_1(\theta) = cw(\theta)/\omega_1 \) is asymptotically \( \chi^2_1 \). For \( d > 1 \), the first-order moment matching can be used, which gives

\[
cw_1(\theta) = \frac{cw(\theta)}{\bar{\omega}},
\]

with \( \bar{\omega} = \sum_{j=1}^{d} \omega_j / d \). A \( \chi^2_d \) approximation is used for the distribution of \( cw_1(\theta) \). A more effective rescaled version of \( cw(\theta) \) is given in Pace et al. (2011).

In principle, the composite likelihood cannot be used directly in Bayes’ theorem as it is not a genuine likelihood. However, Pauli et al. (2011) suggest to combine the composite likelihood \( cL(\theta; y) = \exp\{c_{\ell\theta}(\theta; y)\} \) suitably calibrated, i.e.

\[
cL_c(\theta; y) = cL(\theta; y)^{1/\bar{\omega}},
\]
with a prior $\pi(\theta)$ on $\theta$ in the Bayesian framework to obtain the calibrated composite posterior distribution

$$
\pi_c(\theta|y) \propto \pi(\theta)cL_c(\theta; y).
$$

(7)

The calibration in (6) is necessary in order to adjust the curvature of the composite likelihood (see also Smith & Stephenson, 2009) and allows to approximately recover the asymptotic properties of the full posterior. Examples of (7) are discussed in Pauli et al., 2011; see also Ribatet et al., 2012. A limitation of (7) is that it depends crucially on the calibration factor, whose components are typically cumbersome to compute (see Varin et al., 2011, Section 5.1).

3 ABC with unbiased estimating functions

In the ABC context, the similarity of simulated and observed data is typically measured by means of a distance between some summary statistics, which are in general not sufficient. On the other hand, in order to control the Monte Carlo error, the summary statistics should be as low-dimensional as possible (Fearnhead & Prangle, 2012). In general, the choice of the summary statistics is not straightforward, especially with high-dimensional data and complex model structures.

The approach suggested here uses a suitably rescaled composite score function $cL_\phi(\theta; y)$, evaluated at the observed MCLE $\hat{\theta}_{\text{obs}}$, as a summary of the data. In Section 3.1 we give a motivation for this choice starting from a full computable likelihood function, while the proposed ABC-cs algorithm is discussed in Section 3.2.

3.1 ABC with score functions

Assume that the model belongs to a full exponential family with density

$$
f(y; \varphi) = h(y) \exp\{\varphi^T s(y) - k(\varphi)\},
$$

(8)

where $h(y) > 0$, $\varphi$ is the canonical parameter, $s(y)$ is the $d$-dimensional sufficient statistic, and $k(\varphi)$ is the cumulant generating function of $s(y)$. In this case, the best summary statistic for ABC is the minimal sufficient statistic $s(y)$, which gives the exact posterior for $\epsilon \to 0$ (see, e.g.,
Rubio & Johansen (2013). The following proposition shows that the ABC posterior based on a suitably rescaled score function is exact for \( \epsilon \to 0 \) and also invariant to reparameterizations.

**Proposition 3.1** Let \( \ell(\varphi; y) = \varphi^T s(y) - k(\varphi) \) be the log-likelihood for \( \varphi \) based on model \( \mathbb{X} \), and consider as summary statistic the rescaled score

\[
\eta(y; \varphi_0) = B(\varphi_0)^{-1} \ell_{\varphi}(\varphi_0; y) ,
\]

where \( \ell_{\varphi}(\varphi; y) = \partial \ell(\varphi; y)/\partial \varphi = s(y) - \partial k(\varphi)/\partial \varphi \) and \( B(\varphi) \) is such that \( i(\varphi) = \partial^2 k(\varphi)/\partial \varphi \partial \varphi^T = B(\varphi)B(\varphi)^T \). Then, \( \eta(y; \varphi_0) \) is the best summary statistic in the sense that the ABC posterior based on \( \eta(y; \varphi_0) \) is exact for \( \epsilon \to 0 \) and also invariant to reparameterizations, regardless of the fixed value \( \varphi_0 \).

**Proof** For a fixed value \( \varphi_0 \), the rescaled score \( \eta(y; \varphi_0) \) is a linear transformation of the minimal sufficient statistic \( s(y) \), and thus it is itself minimal sufficient. This proves that the ABC posterior based on \( \eta(y; \varphi_0) \) is exact for \( \epsilon \to 0 \).

Consider the reparametrization \( \theta = \theta(\varphi) \). Let \( \bar{\ell}(\theta) = \ell(\varphi(\theta)) \) and \( \bar{i}(\theta) = \varphi_0^T i(\varphi(\theta)) \varphi_0 \), where \( \varphi_\theta = \partial \varphi(\theta)/\partial \theta \). The rescaled score is \( \bar{\eta}(y; \theta_0) = \bar{B}(\theta_0)^{-1} \bar{\ell}_{\theta}(\theta_0; y) \), with \( \theta_0 = \theta(\varphi_0) \), \( \bar{\ell}_{\theta}(\theta; y) = \partial \bar{\ell}(\theta; y)/\partial \theta \) and \( \bar{B}(\theta) \) such that \( \bar{B}(\theta)\bar{B}(\theta)^T = \bar{i}(\theta) \). Then, since \( \bar{B}(\theta) = \varphi_0^T B(\varphi(\theta)) \) and \( \bar{\ell}_{\theta}(\theta; y) = \varphi_0^T \ell_{\varphi}(\varphi(\theta); y) \), it follows that \( \bar{\eta}(y; \theta_0) = \eta(y; \varphi_0) \). This proves invariance to reparameterizations.\( \blacksquare \)

Proposition 3.1 holds for any value of \( \varphi_0 \). In particular, when \( \varphi_0 \) is the observed value of the maximum likelihood estimate (MLE) at the observed data \( y_{\text{obs}} \), i.e. \( \hat{\varphi}_{\text{obs}} \), we have \( \eta(y_{\text{obs}}; \varphi_{\text{obs}}) = 0 \). This choice of \( \varphi_0 \) is particularly convenient for a general model \( f(y; \theta) \). Indeed, in this case, at least in principle, we could use an alternative representation of \( y \), or equivalently the minimal sufficient statistic based on \( y \), given by \( (\hat{\theta}, a) \), where \( \hat{\theta} \) is the MLE and \( a \) is an ancillary statistic, which means that its distribution does not depend on \( \theta \). Hence, we could replace \( f(y; \theta) \) with \( f(\hat{\theta}, a; \theta) \), and the latter can be factorized as

\[
f(\hat{\theta}, a; \theta) = f(\hat{\theta}|a; \theta)f(a) .
\]

This means that the likelihood for \( \theta \) can be based equivalently on \( f(y; \theta) \) or \( f(\hat{\theta}|a; \theta) \). Unfortunately, it may not be easy in general to find \( f(\hat{\theta}|a; \theta) \). On the other hand, it is possible
to approximate such density through a tangent exponential model at (and near) the fixed value \( y_{obs} \) (Fraser & Reid, 1995; Reid, 2003, Sect. 3.2). Denoting by \( \ell(\theta; y_{obs}) \) the observed log-likelihood, the approximation to the log-likelihood based on the tangent exponential is

\[
\ell_{TE}(\theta; y) = \ell(\theta; y_{obs}) - \ell(\hat{\theta}_{obs}; y_{obs}) + \{\varphi(\theta) - \varphi(\hat{\theta}_{obs})\}^T s(y),
\]

where \( \hat{\theta}_{obs} \) is the MLE at the observed data point \( y_{obs} \), \( s(y) = \partial\ell(\theta; y)/\partial\theta|_{\theta=\hat{\theta}_{obs}} = \ell(\hat{\theta}_{obs}; y) \), and \( \varphi(\theta) = \varphi(\theta; y_{obs}) \) is a one-to-one reparameterization dependent on the observed data \( y_{obs} \) (see also Brazzale et al., 2007, Sect. 8.4.2). The tangent exponential model is a local exponential family model with sufficient statistic \( s(y) \) and canonical parameter \( \varphi \). It has the same log-likelihood function as the original model at the fixed point \( y_{obs} \), where it also has the same first derivative with respect to \( y \).

From Proposition 3.1 the best summary statistic for ABC for the tangent exponential model (9) is the rescaled score, where the score is given by

\[
\ell_{TE}(\theta; y) = \ell(\theta; y_{obs}) + \varphi_{\theta} s(y) .
\]

For \( \theta = \hat{\theta}_{obs} \), (10) reduces to \( \varphi_{\theta}(\hat{\theta}_{obs})\ell(\hat{\theta}_{obs}; y) \), i.e. to a linear transformation of the score of the original model. Rescaling (10) then provides invariance to reparameterization, as in Proposition 3.1. This motivates the use of the score function evaluated at \( \hat{\theta}_{obs} \) as an approximate optimal summary statistic in ABC for a general model.

**Example 1: normal parabola.** Let \( y = (y_1, \ldots, y_n) \) be a random sample from the normal distribution \( N(\theta, \theta^2) \), with \( \theta > 0 \). The log-likelihood is

\[
\ell(\theta; y) = \frac{1}{\theta} \sum_{i=1}^{n} y_i - \frac{1}{2\theta^2} \sum_{i=1}^{n} y_i^2 - n \log \theta ,
\]

where \( t(y) = (\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} y_i^2) \) is the two-dimensional minimal sufficient statistic. The score function is \( \ell_{\theta}(\theta; y) = -\theta^{-2} \sum_{i=1}^{n} y_i + \theta^{-3} \sum_{i=1}^{n} y_i^2 - n\theta^{-1} \), which implies that \( \hat{\theta} \) is the positive solution of a quadratic equation. The expected information is \( i(\theta) = 3n/\theta^2 \), and the rescaled score is \( \eta(y; \hat{\theta}_{obs}) = \hat{\theta}_{obs} \ell_{\theta}(\hat{\theta}_{obs}; y)/\sqrt{3n} \).

As an illustration we use a sample of size \( n = 50 \) generated from the model, with \( \theta = 5 \) and with a uniform prior in (0,15). We consider three instances of the ABC Algorithm [1]
with distance $\rho(v, w) = ||v - w||_1$ and with summary statistics given, respectively, by $t(y)$, $\eta(y; \hat{\theta}_{\text{obs}})$, and also a one-to-one transformation of the minimal sufficient statistic $t(y)$, that is $t_1(y) = (\bar{y}, \sqrt{s^2})$, i.e. the sample mean and standard deviation. In all three cases we used the same sample of $10^7$ values generated from the prior and in each case we chose the threshold $\epsilon$ as the quantile of level 0.1\% of the observed distances, thus accepting $10^4$ values. These $\epsilon$ values are respectively 31.264, 0.02 and 0.237.

Figure 1: Normal parabola. In all panels the solid line corresponds to the exact posterior, while the dashed lines correspond to ABC approximations using $t(y)$ (left panel), $t_1(y)$ (central panel), and $\eta(y; \hat{\theta}_{\text{obs}})$ (right panel).

Figure 1 shows the three approximations compared with the exact posterior. The two versions of the ABC with the minimal sufficient statistics gave quite different results, with the one with $t(y)$ leading to the worst accuracy. This is likely due to the large value of $\epsilon$ (31.264). We note that only three of the $10^7$ proposed values of $\theta$ would have been accepted with $\epsilon = 1$, thus making the ABC algorithm with $t(y)$ impractical. On the other hand, the ABC with the one-dimensional summary statistic $\eta(y; \hat{\theta}_{\text{obs}})$, which is not sufficient for this model, gave an approximation to the posterior with accuracy comparable with ABC with the minimal sufficient statistic $t_1(y)$.

**Remark 1.** From the point of view of the likelihood principle, the different performances of the ABC algorithm with the two versions of the minimal sufficient statistic in Example 1 is unpleasant. Indeed, $t(y)$ and $t_1(y)$ lead to the same likelihood and posterior functions
but the two ABC approximations could be remarkably different, as in the example above. On the contrary, since the likelihood and the score functions are not affected by one-to-one transformations of the data, or of the minimal sufficient statistic, ABC with $\eta(y; \hat{\theta}^{\text{obs}})$ is invariant with respect to such transformations.

Remark 2. Although the choice of the distance function in the ABC algorithm $\rho(\cdot, \cdot)$ is arbitrary, when considering the Euclidean distance we have

$$
\rho \left( \eta(y; \hat{\theta}^{\text{obs}}), \eta(y^{\text{obs}}; \hat{\theta}^{\text{obs}}) \right) = ||\eta(y; \hat{\theta}^{\text{obs}})||_{1/2}^2 = \left\{ \ell_{\theta}(\hat{\theta}^{\text{obs}}; y)^T i(\hat{\theta}^{\text{obs}})^{-1} \ell_{\theta}(\hat{\theta}^{\text{obs}}; y) \right\}^{1/2},
$$

which is the score test statistic computed in $\hat{\theta}^{\text{obs}}$, based on data $y$.

Despite the good properties of ABC with the score function, unfortunately in typical applications of the ABC method the likelihood function is intractable, and therefore the same holds for the score function. This motivates the extension to composite likelihoods proposed in the next section.

### 3.2 ABC with composite score function

When dealing with complex models, possible surrogates of the unavailable full likelihood are given by composite likelihoods. Analogously to what seen in the previous section for a full likelihood, we propose the rescaled composite score function as a summary statistic in ABC. This defines an algorithm, called ABC-cs. In terms of the ABC Algorithm\textsuperscript{[i]} ABC-cs replaces the matching condition

$$
\rho(\eta(y), \eta(y^{\text{obs}})) \leq \epsilon,
$$

with

$$
\rho \left( \eta_c(\hat{\theta}^{\text{obs}}; y), \eta_c(\hat{\theta}^{\text{obs}}; y^{\text{obs}}) \right) \leq \epsilon, \tag{11}
$$

where $\hat{\theta}^{\text{obs}}$ is the MLE computed with $y^{\text{obs}}$ and

$$
\eta_c(\hat{\theta}^{\text{obs}}; y) = B_c(\hat{\theta}^{\text{obs}})^{-1} c\ell_{\theta}(\hat{\theta}^{\text{obs}}; y) \tag{12}
$$

is the rescaled composite score, with $B_c(\theta)$ such that $J(\theta) = B_c(\theta)B_c(\theta)^T$. Since $c\ell_{\theta}(\hat{\theta}^{\text{obs}}; y^{\text{obs}}) = 0$, in \textsuperscript{[i]} we only need to evaluate $\eta_c(\hat{\theta}^{\text{obs}}; y)$. 

12
An advantage of ABC-cs is that the rescaled composite score statistic has the same dimension as \( \theta \), hence the complexity of the method is linear in the number of parameters. Moreover, since the score statistic is obtained from the composite log-likelihood by just taking the first derivative, it is easily computed, especially when it is analytically available. An apparent drawback of (12) is the implicit dependence of the ABC-cs algorithm on \( J(\theta) \). However, only \( J(\tilde{\theta}_{\text{obs}}) \) is needed, and this quantity can be easily approximated with a preliminary Monte Carlo simulation with few hundreds replications (Cattelan & Sartori, 2014).

The following theorem shows that the proposed ABC-cs algorithm gives a valid approximation to the posterior distribution even if the rescaled composite score function (12) does not satisfies the information identity, as a full score function.

**Theorem 3.2** ABC-cs with the rescaled score statistic \( \eta_c(\tilde{\theta}_{\text{obs}}; y) \) leads to an approximate posterior distribution with the correct curvature and is also invariant to reparameterizations.

**Proof** In order to recover the information identity, and thus the correct curvature, it is necessary to consider the adjusted composite score function (see, e.g., Pace & Salvan, 1997, Chap. 4)

\[
g(\theta; y) = H(\theta)J(\theta)^{-1}c\ell_\theta(\theta; y) = A(\theta)c\ell_\theta(\theta; y)
\]

Indeed, for \( g(\theta; y) \) we have

\[
J_g(\theta) = \text{var}_\theta\{g(\theta; Y)\} = A(\theta)\text{var}_\theta\{c\ell_\theta(\theta; Y)\}A(\theta)^T = G(\theta)
\]

and

\[
H_g(\theta) = E_\theta \left\{ -\frac{\partial}{\partial \theta^T} g(\theta; Y) \right\}
\]

\[
= - \left\{ \frac{\partial}{\partial \theta^T} A(\theta) \right\} E_\theta\{c\ell_\theta(\theta; Y)\} - A(\theta)E_\theta \left\{ \frac{\partial}{\partial \theta^T} c\ell_\theta(\theta; Y) \right\} = G(\theta).
\]

Since \( H_g(\theta) = J_g(\theta) = G(\theta) \), the adjusted composite score \( g(\theta; y) \) satisfies the information identity as a proper score function and, since \( A(\theta) \neq 0 \), it leads to the same estimator \( \tilde{\theta} \) of \( c\ell_\theta(\theta; y) = 0 \).

The ABC-cs algorithm should be based on the rescaled version of \( g(\theta; y) \), given by

\[
\eta_g(\tilde{\theta}_{\text{obs}}; y) = B_g(\tilde{\theta}_{\text{obs}})^{-1} g(\tilde{\theta}_{\text{obs}}; y)
\]
where $B_g(\theta) = H(\theta)\{B_c(\theta)^T\}^{-1}$. Indeed,

$$G(\theta) = H(\theta)J(\theta)^{-1}H(\theta) = H(\theta)\{B_c(\theta)^T\}^{-1}B_c(\theta)^{-1}H(\theta).$$

However, it is straightforward to see that

$$\eta_g(\theta; y) = B_g(\theta)^{-1}g(\theta; y) = B_c(\theta)^T H(\theta)^{-1}H(\theta)J(\theta)^{-1}\ell(\theta; y) = \eta_c(\theta; y).$$

This proves that the use of $\eta_c(\tilde{\theta}_{obs}; y)$ as a summary statistic for ABC leads to an approximate posterior with the correct curvature (see Pauli et al., 2011).

The proof of invariance to reparameterization follows the same steps as in Proposition 3.1.

As a final remark, note that even in this case, the squared Euclidean distance gives the composite score test statistic computed in $\tilde{\theta}_{obs}$, based on data $y$.

4 Examples

In the following examples we use composite marginal likelihood functions (Cox & Reid, 2004), although different model structures might lead to different choices of suitable composite likelihoods. An advantage of the ABC-cs approach is that it gives (with a small additional effort) the Godambe information $G(\tilde{\theta}_{obs})$, which we use throughout the examples as a precision matrix both for ABC and ABCcs with importance sampling. Notice that ABC with MCMC or sequential Monte Carlo methods, requires a similar precision matrix, which in practice is unknown, and must be estimated by considering several preliminary runs of ABC. All the examples considered here can be replicated in R by using the package ABCcs, available at http://homes.stat.unipd.it/erlisruli/en/content/publications.

Example 2: equi-correlated normal model

This example, considered in Pace et al. (2011) among others, focuses on Bayesian inference based on the pairwise log-likelihood for the parameters of an equi-correlated multivariate normal distribution with mean vector $\mu$, and covariance matrix $\Sigma_{rs} = \rho \sigma^2$, for $r \neq s$, and
$\Sigma_{rr} = \sigma^2$, $r, s = 1, \ldots, q$. For this model, $\hat{\theta}$ is full efficient, the sufficient statistic is three-dimensional and is the same for both the full and pairwise likelihoods (Pace et al., 2011). The pairwise log-likelihood (3) for $\theta = (\mu, \sigma^2, \rho)$ is

$$p\ell(\theta; y) = -\frac{nq(q - 1)}{2} \log \sigma^2 - \frac{nq(q - 1)}{4} \log(1 - \rho^2) - \frac{q - 1 + \rho}{2\sigma^2(1 - \rho^2)} SS_W$$

$$- \frac{q(q - 1)SS_B + nq(q - 1)(\bar{y} - \mu)^2}{2\sigma^2(1 + \rho)},$$

where $SS_W = \sum_{i=1}^{n} \sum_{r=1}^{q} (y_{ir} - \bar{y})^2$, $SS_B = \sum_{i=1}^{n} (\bar{y}_i - \bar{y})^2$, $\bar{y}_i = \sum_{r=1}^{q} y_{ir}/q$ and $\bar{y} = \sum_{i=1}^{n} \sum_{r=1}^{q} y_{ir}/(nq)$. For the expression of the score function see Pace et al. (2011, p. 145).

We assume the components of the parameter $\omega = (\mu, \tau, \kappa)$, with $\tau = \log \sigma^2$ and $\kappa = \logit\left(\frac{\rho(q - 1) + 1}{q}\right)$, independent, with $N(0, 100)$ marginal prior distributions.

A sample of $n = 30$ is drawn from the model with $q = 50$, $\mu = 0$, $\sigma^2 = 1$ and $\rho = 0.5$. For ABC the summary statistic is $(\bar{y}, \sqrt{SS_B}, \sqrt{SS_W})$, while for ABC-cs the summary statistic is given by (12). The simulation from the ABC and ABC-cs posteriors is performed with importance sampling, where the importance function is the multivariate $t$-student distribution with 5 degrees of freedom, centred at $\tilde{\theta}_{\text{obs}}$ and with scale matrix equal to five times $G(\tilde{\theta}_{\text{obs}})^{-1}$. We consider $10^3$ final samples obtained with $\epsilon$ fixed to the 0.1% quantile of the observed distances. To get rid of the importance weights we consider re-sampling with replacement.

Results are compared also with the pairwise posterior

$$\pi_{pl}(\theta|y) \propto \pi(\theta) \exp\{p\ell(\theta; y)\},$$

with the pairwise posterior (11) based on the calibrated pairwise likelihood and with the posterior distribution based on the full likelihood, approximated by a random walk Metropolis.

The box-plots of the several marginal posterior approximations are shown in Figure 2. The figure highlights several interesting features. As is well known, the posterior (13) can be wrongly too concentrated (see also Pauli et al., 2011; Smith & Stephenson, 2009; Ribatet et al., 2012), whereas the calibrated pairwise posterior (7) may be the opposite. Indeed, while the marginal calibrated pairwise posteriors of $\mu$ and $\tau$ are quite similar to the full posterior (MCMC), the marginal calibrated pairwise posterior of $\kappa$ is slightly narrower that the corresponding marginal based on the full likelihood.
On the other hand, ABC-cs and ABC marginal posteriors are all quite similar to the full posterior. This is not surprising, since the model is a full exponential family of order three and ABC uses exactly the sufficient statistic as summary statistic. Moreover, even the pairwise likelihood has exponential form, with the same sufficient statistic. This implies that the pairwise score function is proportional to the score function of the full model (Kenne Pagui, 2013, Theorem 1, pag. 14) and the latter would lead again to the sufficient statistic (see Section 3.1).

We also compare the posterior means of the full, ABC and ABC-cs posteriors in a simulation study over 100 Monte Carlo trials. The data are generated from the model with $\mu = 0$, $\sigma^2 = 1$, $\rho = \{0.2, 0.5, 0.9\}$. At each simulated dataset, the ABC, the ABC-cs and the full posteriors are approximated as above. From the simulations (see Figure 3), we notice that ABC and ABC-cs posterior means are quite similar to the full posterior mean, as expected from Proposition 3.1. Whereas the mean of the calibrated pairwise posterior can perform poorly.

**Example 3: multilevel probit**

The pairwise likelihood is particularly useful for modelling correlated binary outcomes, as discussed in Le Cessie & van Houwelingen (1994). This kind of data arise, e.g., in the context of repeated measurements on the same individual. Standard likelihood analysis in these contexts may be difficult because it involves multivariate integrals whose dimension equals
Figure 3: Equi-correlated normal model. Simulation study based on 100 Monte Carlo trials, with $\mu = 0$, $\sigma = 1$ ($\tau = 0$). The dashed horizontal lines represent the true parameter values.
the cluster sizes.

Let us focus on a multilevel probit model with constant cluster sizes. In particular, let $S_i$ be a latent and unobserved $q$-variate normal with mean $\gamma_i = X_i \beta$, with $\beta$ vector of unknown regression coefficients, $X_i$ design matrix for unit $i$, and covariance matrix $\Sigma$, with 

$$\Sigma_{hh} = 1 + \sigma^2, \quad \Sigma_{hk} = \sigma^2, \quad h \neq k, \quad i = 1, \ldots, n.$$  

Then, the observed data $Y_{ih}$ is equal to 1 if $S_{ih} > 0$, and 0 otherwise, for $h = 1, \ldots, q$, $i = 1, \ldots, n$.

The full likelihood is cumbersome since it entails calculation of multiple integrals of a $q$-variate multivariate normal distribution. On the other hand, the pairwise log-likelihood is

$$p\ell(\beta, \sigma^2; y) = \sum_{i=1}^{n} \sum_{h=1}^{q-1} \sum_{k=h+1}^{q} \log \Pr(Y_{ih} = y_{ih}, Y_{ik} = y_{ik}; \beta, \rho), \quad y_{ih}, y_{ik} \in \{0, 1\},$$

where, for instance, $\Pr(Y_{ih} = 1, Y_{ik} = 1; \beta, \rho) = \Phi_2(\gamma_{ih}, \gamma_{ik}; \rho)$ is the standard bivariate normal distribution, with correlation $\rho = \sigma^2/(1 + \sigma^2)$ and with $\gamma_{ih} = x_{ih}\beta/\sqrt{1 + \sigma^2}$ the $h$ component of $\gamma_i$ ($i = 1, \ldots, n; h, k = 1, \ldots, q$).

As an example, we consider data generated with $\beta_0 = 0.5$, $\beta_1 = 1.5$, $\sigma^2 = 1$, $n = 30$ and $q = 10$, where $\beta_0$ is the intercept and $\beta_1$ the coefficient of a covariate generated from $U(-1, 1)$. For the parameter $\theta = (\beta_0, \beta_1, \log \sigma^2)$ a trivariate normal prior with independent components $N(0, 100)$ is assumed. For ABC we take the counts at each time point $h$ ($h = 1, \ldots, q$), as summary statistic which is $q$-dimensional. Hence, the absolute norm among the statistics is $\sum_{h=1}^{q} |\sum_{i=1}^{n} (y_{ih}^{\text{obs}} - y_{ih})|$. Moreover, considering the whole data matrix $y^{\text{obs}}$ as summary statistic would perform worse. For ABC-cs, we consider the rescaled pairwise score, evaluated at $\tilde{\theta}^{\text{obs}}$.

The matrices $J(\tilde{\theta}^{\text{obs}})$ and $H(\tilde{\theta}^{\text{obs}})$ were computed by simulation with 1000 datasets taken from the model with $\theta = \tilde{\theta}^{\text{obs}}$. The computation of the bivariate normal integrals is performed by calling Fortran routines available from Alan Genz’s website (http://www.math.wsu.edu/faculty/genz/homepage). We consider $10^3$ final samples drawn from the ABC and ABC-cs posteriors after fixing $\epsilon$ to the 0.1% quantile of the observed distances. The sampling is done via importance sampling, with a multivariate $t$-student importance density, with 5 degrees of freedom, centred at $\tilde{\theta}^{\text{obs}}$ and with scale matrix equal to five times $G(\tilde{\theta}^{\text{obs}})^{-1}$.

If we treat the calibrated pairwise posterior as a benchmark, than we notice that again the ABCcs posterior tends to work better that the ABC approximation. On the other hand
the non calibrated pairwise posterior is overly concentrated.

Figure 4: Multilevel probit. ABC-cs posterior compared with the ABC, the pairwise (pair), and the adjusted pairwise posteriors.

A simulation study is conducted over 100 Monte Carlo samples, where the covariate is simulated as previously, with \( \beta_0 = 0.5 \), \( \beta_1 = 1.5 \), \( n = 30 \), \( q = 10 \) and \( \sigma^2 = \{1, 2.5, 4\} \). For each simulated dataset, we consider the mean of \( 10^3 \) final samples drawn from the four posteriors considered previously. The simulation algorithm is the same as above.

From the simulations shown in Figure 5 it is evident that the ABC mean performs very poorly also in repeated sampling. On the other hand, the ABC-cs mean is more accurate, even with respect to the mean of the adjusted pairwise posterior. In this case also the non calibrated pairwise posterior mean is performing reasonably. Nevertheless, as shown in Figure 4, this posterior is overly precise.

**Example 4: MA(2) process**

Consider an MA\((p)\) process, defined as

\[
Y_t = u_t + \sum_{i=1}^{p} \theta_i u_{t-i},
\]

where \( u_t, \ t = 1, \ldots, q \), are independent \( N(\mu, \sigma^2) \), and \( \theta_i, \ i = 1, \ldots, p \), must satisfy the identifiability conditions, namely that the roots of the polynomial

\[
Q(x) = 1 - \sum_{i=1}^{p} \theta_i x^i
\]
Figure 5: Multilevel probit. Simulations based on 100 Monte Carlo trials, with $\beta_0 = 0.5$, $\beta_1 = 1$. 

(b) $\sigma^2 = 1$ ($\log \sigma^2 = 0$)

(a) $\sigma^2 = 2.5$ ($\log \sigma^2 = 0.92$)

(c) $\sigma^2 = 4$ ($\log \sigma^2 = 1.39$)
are all outside of the unit circle in the complex plane. This stochastic process is typically used for time series modelling.

The likelihood of the MA($p$) model, obtained by integrating out the random components $u_t$ (see, e.g., Hamilton, 1994), involves inversions of ($q \times q$) covariance matrices, which for large $p$ and $q$ may be computationally challenging owing to the matrix inversions. A better approach is to resort to the Kalman filter (see Hamilton, 1994, Ch. 13). However, as shown by Marin et al. (2012), the ABC algorithm works reasonably well in this example, so it is instructive to compare it with ABC-cs based on the composite likelihood.

We focus on the MA(2) model. As in Marin et al. (2012), we assume $\mu = 0$, $\sigma^2 = 1$, and the prior for $\theta = (\theta_1, \theta_2)$ is assumed uniform in the parameter space, i.e. the triangle $-2 < \theta_1 < 2$, $\theta_1 + \theta_2 > -1$, $\theta_1 - \theta_2 < 1$.

We use as summary statistics for ABC the first three autocovariances

$$\tau_j = \sum_{t=j+1}^{q} y_t y_{t-j}, \quad j = 0, 1, 2.$$ 

In this example, given the model structure (see e.g. Hamilton, 1994, p. 130), we use a triplewise log-likelihood of the form

$$c\ell(\theta; y) = \sum_{t=1}^{q-2} \log f(y_t, y_{t+1}, y_{t+2}; \theta).$$

As in Marin et al. (2012), we draw $q = 100$ values from the MA(2) model, with parameters $(\theta_1, \theta_2) = (0.6, 0.2)$. For the ABC-cs posterior the summary statistic is the rescaled triplewise score, evaluated at $\tilde{\theta}^{obs}$. A sample of $10^3$ final values is drawn from the ABC and ABC-cs posteriors with Algorithm 1, and $\epsilon = 0.1\%$ quantile of the absolute distances among the statistics. For illustration purposes the ABC and ABC-cs posteriors are compared also with the full posterior approximated with a random walk Metropolis algorithm. From the posteriors, shown in Figure 6, we notice that the ABC-cs approximation tends to be slightly better than ABC.

A simulation study is performed, with 100 Monte Carlo samples drawn from the model with the parameter $(\theta_1, \theta_2) = (0.6, 0.2)$. For each simulated dataset, we run ABC and ABC-cs
Figure 6: MA(2) model. Top panel: comparison of the level sets (in black) of the posterior distribution with a smoothed scatter plot of the ABC posterior and ABC-cs (crosses) and with box-plots of the ABC posterior (blue) and ABC-cs. Bottom-left (bottom-right) panel: histogram of the marginal posterior of $\theta_1$ ($\theta_2$), compared with ABC (continued) and ABC-cs (dashed).

with $10^3$ final samples drawn as above. Over this final draws the average is taken and it is compared also with the average of the MCMC samples taken form the full posterior. The simulation results are plotted in Figure 7. Both ABC methods give reasonable results when compared to the full posterior, with ABC-cs being overall preferable to ordinary ABC.
Figure 7: MA(2) model. Comparisons of the full, ABC and ABC-cs posterior means in 100 Monte Carlo trials, with $(\theta_1, \theta_2) = (0.6, 0.2)$ (horizontal lines).

5 Discussion

A new procedure for constructing summary statistics for ABC is proposed, which is based on score or composite score functions. An advantage of the proposed method is that, by construction, the summary statistics automatically incorporate relevant features of the complex model, and its dimension is the same as the number of parameters. Moreover, no post processing tasks are required, nor pilot runs or ad hoc summaries of the data. The proposed approach can be used also within more elaborate Monte Carlo algorithms, such as MCMC, or sequential Monte Carlo methods.

The success of the ABC-cs procedure depends on how good is the composite likelihood as an approximation for the full model, given the observed data. In complex models, composite likelihoods are ideal inferential tools for deriving useful parameter estimates. Although in the examples we focused mainly on composite marginal likelihoods, this is only a special case of the general class of composite likelihoods. Indeed, there exists a wide range of possibilities for constructing composite likelihoods, and the choice depends on the structure and complexity of the model at hand. There is a rich and growing literature on this topic (Varin et al., 2011), which we believe may be fruitfully used in ABC applications.

Finally, we note that we used the composite likelihood as a natural basis to construct
a suitable unbiased estimating function in complex models. However, the proposed ABC algorithm works with any unbiased estimating function, such as for instance those used in the robust literature (see, e.g., [Huber & Ronchetti, 2009]).

References

BARTHÉLÉMY, S. & CHOPIN, N. (2014). Expectation propagation for likelihood-free inference. *Journal of the American Statistical Association* **109**, 315–333.

BEAUMONT, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics* **41**, 379–406.

BEAUMONT, M. A., CORNUET, J.-M., MARIN, J.-M. & ROBERT, C. P. (2009). Adaptive approximate Bayesian computation. *Biometrika* **96**, 983–990.

BRAZZALE, A. R., DAVISON, A. C. & REID, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge: Cambridge University Press.

CATTELAN, M. & SARTORI, N. (2014). Empirical and simulated adjustments of composite likelihood ratio statistics. *arXiv* arXiv:1403.7093v1.

COX, D. R. & REID, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729–737.

DIGGLE, P. J. & GRATTON, R. J. (1984). Monte Carlo Methods of inference for implicit statistical models (with Discussion). *Journal of the Royal Statistical Society: Series B* **46**, 193–227.

DROVANDI, C. C., PETTITT, A. N. & FADDY, M. J. (2011). Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society: Series C* **60**, 317–337.

ERHARDT, R. J. & SMITH, R. L. (2012). Approximate Bayesian computing for spatial extremes. *Computational Statistics & Data Analysis* **56**, 1468–1481.
Fearnhead, P. & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation (with Discussion). *Journal of the Royal Statistical Society: Series B* **74**, 419–474.

Fraser, D. A. S. & Reid, N. (1995). Ancillaries and third order significance. *Utilitas Mathematica* **47**, 33–53.

Gourieroux, C., Monfort, A. & Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics* **8**, S85–S118.

Hamilton, J. D. (1994). *Time series analysis*. Princeton: Princeton University Press.

Heggland, K. & Frigessi, A. (2004). Estimating functions in indirect inference. *Journal of the Royal Statistical Society: Series B* **66**, 447–462.

Hjort, N. L. & Varin, C. (2008). ML, PL, QL in Markov chain models. *Scandinavian Journal of Statistics* **35**, 64–82.

Huber, P. J. & Ronchetti, E. M. (2009). *Robust Statistics*. Hoboken, New Jersey: Wiley.

Kenne Pagui, E. C. (2013). *Combined Composite Likelihoods*. Ph.D. thesis, Department of Statistical Science, University of Padova.

Larribe, F. & Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statistica Sinica* **21**, 43.

Le Cessie, S. & van Houwelingen, J. C. (1994). Logistic regression for correlated binary data. *Journal of the Royal Statistical Society: Series C* **43**, 95–108.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 220–239.

Lindsay, B. G., Yi, G. Y. & Sun, J. (2011). Issues and strategies in the selection of composite likelihoods. *Statistica Sinica* **21**, 71.
MARIN, J.-M., PUDLO, P., ROBERT, C. P. & RYDER, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing* **22**, 1167–1180.

MARJORAM, P., MOLITOR, J., PLAGNOL, V. & TAVARÉ, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences* **100**, 15324–15328.

MENGERSEN, K. L., PUDLO, P. & ROBERT, C. P. (2013). Bayesian computation via empirical likelihood. *Proceedings of the National Academy of Sciences* **110**, 1321–1326.

MINKA, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc.

MOLENBERGHS, G. & VERBEKE, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.

PACE, L. & SALVAN, A. (1997). *Principles of Statistical Inference*. Singapore: World Scientific.

PACE, L., SALVAN, A. & SARTORI, N. (2011). Adjusting composite likelihood ratio statistics. *Statistica Sinica* **21**, 129–148.

PAULI, F., RACUGNO, W. & VENTURA, L. (2011). Bayesian composite marginal likelihoods. *Statistica Sinica* **21**, 149–164.

REID, N. (2003). Asymptotics and the theory of inference. *The Annals of Statistics* **31**, 1695–1731.

RIBATET, M., COOLEY, D. & DAVISON, A. C. (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica* **22**, 813–845.

RUBIO, F. J. & JOHANSEN, A. M. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics* **7**, 1632–1654.
Sisson, S., Fan, Y. & Tanaka, M. (2007). Sequential Monte Carlo without likelihoods. 
*Proceedings of the National Academy of Sciences of the United States of America* **104**, 1760–1765.

Sisson, S., Fan, Y. & Tanaka, M. (2009). Sequential Monte Carlo without likelihoods: 
Errata. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 16889.

Smith, E. L. & Stephenson, A. G. (2009). An extended Gaussian max-stable process 
model for spatial extremes. *Journal of Statistical Planning and Inference* **139**, 1266–1275.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis* **92**, 1–28.

Varin, C., Reid, N. & Firth, D. (2011). An overview of composite likelihood methods. 
*Statistica Sinica* **21**, 5–42.