Robust approximate Bayesian inference with an application to linear mixed models

Erlis Ruli, Nicola Sartori and Laura Ventura

Department of Statistical Sciences, University of Padova, Italy
ruli@stat.unipd.it, sartori@stat.unipd.it, ventura@stat.unipd.it

June 7, 2017

Abstract

We discuss an approach for deriving robust posterior distributions from robust M-estimating functions using Approximate Bayesian Computation (ABC) methods. The method is formally motivated by the use of unbiased estimating functions as summary statistics in ABC and the theoretical properties of the robust posterior distributions are discussed. The motivation for the development of the method is given by an application of linear mixed models in a clinical study with repeated measurements. Both the application and simulation results show the usefulness of the method. An R implementation is also provided through the robustBLME package available on CRAN.

Keywords: Empirical likelihood; influence function; likelihood-free inference; M-estimators; quasi-likelihood; robustness; summary statistic; unbiased estimating function.

1 Introduction

The normal assumption is the basis of statistical analyses in several fields, such as medicine and health sciences. Indeed, under this assumption, standard parametric estimation and
testing procedures are simple and most efficient. However, it is well known that these procedures are not robust when the normal distribution is just an approximate parametric model or in the presence of outliers in the observed data. A challenging example, characterised by the possible presence of outliers, is given by the study of Tramentozzi et al. (2016) discussed in Section 2, which focuses on the role of a protein in the production of immunoglobulines and cytokines. In particular, the study concerns repeated measures of moderate sample size, for which it is shown that the classical Bayesian analysis through the normal Linear Mixed Model (LMM) is strongly influenced by outliers. In this case, it may be preferable to base inference on procedures that are more resistant, which specifically take into account the fact that the assumed models is only approximate. In order to produce statistical procedures that are stable with respect to small changes in the data or to small model departures, robust statistical methods can be considered.

The concept of robustness has been widely discussed in the frequentist literature; see, for instance, Hampel et al. (1986), Tsou and Royall (1995), Lavine (1995) and Markatou et al. (1998). On the other hand, Bayesian robustness literature is also well developed, most of which is focused on sensitivity with respect to the prior distribution; we mention here the overviews in Berger (1994), Ríos Insua and Ruggeri (2000) and Ruggeri et al. (2005). Recently, Bayesian robustness with respect to model misspecification have attracted considerable attention. For instance, Lazar (2003), Greco et al. (2008) and Ventura et al. (2010) discuss approaches based on robust pseudo-likelihood functions, i.e. the quasi-likelihood and the empirical likelihood, derived from robust M-estimation functions, as replacement of the genuine likelihood in Bayes’ formula. Lewis et al. (2014) discuss an approach for building posterior distributions from robust M-estimators using constrained Markov chain Monte Carlo (MCMC) methods. Approaches based on tilted likelihoods can be found in Miller and Dunson (2015), Watson and Holmes (2016) and Grünwald and van Ommen (2014), among others. Finally, approaches based on model embedding through heavy-tailed distributions or on weighted likelihoods are discussed by Andrade and O’Hagan (2006) and Agostinelli and Greco (2013), respectively.

However, the aforementioned approaches are not without limitations. In particular, robust posterior distributions derived from robust estimating functions have two drawbacks: the
empirical likelihood is not computable for very small sample sizes and for moderate sample sizes appears to have always heavy tails; posterior distributions based on the quasi-likelihood can be easily obtained only for scalar parameters. The restricted likelihood approach of Lewis et al. (2014), as well as the previous approaches based on estimating equations can be computationally cumbersome with robust $M$-estimating functions in the context of LMM (see Sect. 5). The solution of embedding in a larger structure has the cost of eliciting a prior distribution for the extra parameters introduced in the analysis. Moreover, the statistical procedures derived under an embedded model are not necessarily robust in a broad sense, since the larger model may still be too restricted in the space of all distributions. Finally, the tilted and the weighted likelihood approaches refer to concepts of robustness that are not directly related to the one considered in this paper.

Concretely, we focus on the robustness approach based on the influence function ($IF$), extensively discussed in book-length treatments by Huber (1981), Hampel et al. (1986) and Huber and Ronchetti (2009), and on the derivation of robust posterior distributions that condition on, generally insufficient, statistics with bounded $IF$. Such a conditioning is performed by appealing to the well-developed frequentist robust $M$-estimation theory, i.e. unbiased $M$-estimating equations with bounded $IF$, which provides estimators that target the parameter of interest and satisfy the usual asymptotical properties such as unbiasedness and normality. $M$-estimating functions are general unbiased estimating functions that include as special case the score function and scoring rules (see, e.g., Dawid and Musio, 2014, and references therein). To overcome the previous difficulties with current approaches to Bayesian robustness based on unbiased estimating equations, we propose an alternative method based on Approximate Bayesian Computation (ABC) (see, e.g., Pritchard et al., 1999). The typical motivation for the use of ABC is the intractability of the likelihood. Here, instead, we couple ABC with robust unbiased $M$-estimating functions. The idea can be seen as an extension of the results of Ruli et al. (2016) to Bayesian robustness with respect to misspecification. The method is easy to implement and computationally efficient, even when the $M$-estimating functions are cumbersome to evaluate for different parameter values. Theoretical properties, implementation details and simulations results are discussed.

Section 2 describes a motivating dataset on the glucose regulated protein94 (GRP94).
Section 3 sets the necessary background. Section 4 describes the proposed method and its theoretical properties. Section 5 investigates the properties of the proposed method in the context of LMM through simulations and applies it to the GRP94 dataset. Concluding remarks are given in Section 6.

2 The GRP94 dataset

The GRP94 dataset (Tramentozzi et al., 2016) concerns the measurement of glucose-regulated protein94 in plasma or other biological fluids and the study of its role as a tumour antigen, i.e. its ability to alter the production of immunoglobines (IgGs) and inflammatory cytokines in the peripheral blood mononuclear cells (PBMCs) of tumour patients. The study involved 28 patients admitted to the division of General Surgery of the Civil Hospital of Padova for ablation of primary, solid cancer of the gastro-intestinal tract. For each patient, gender, age (expressed in years), type and stage of tumour (ordinal scales of four levels) are given. Patient with ID 15 was removed from the study afterwards, for clinical reasons. Patients’ plasma and PBMCs were challenged with GRP94 complexes and the level of IgG and of the cytokines: interferonγ (IFNγ), interleukin 6 (IL-6), interleukin 10 (IL-10) and tumour necrosis factor α (TNFα) were measured. Owing to time and cost constraints, for patients IDs 17, 27 and 28 only the IgG were measured. The following five treatments were considered: GRP94 at the dose of either 10 ng/ml or 100 ng/ml, GRP94 in complex with IgG (GRP94+IgG) at the doses 10 ng/ml or 100 ng/ml and IgG at the dose 100 ng/ml. Finally, baseline measurements of IgG and of the aforementioned cytokines were taken from untreated PMBCs. Although fresh patient’s plasma and PMBCs are taken for each treatment and patient, the resulting measures are likely to be correlated since plasma and PMBCs are taken from the same patient. Hence, a LMM is a suitable model for this data. Using paired Mann-Whitney tests Tramentozzi et al. (2016) showed that GRP94 in complex with IgG at the higher dose can significantly inhibit the production of IgG and stimulate the secretion of IL-6 and TNFα from PBMCs of cancer patients. In addition, some of the differences between treatments were significant for a specific gender; see Tramentozzi et al. (2016) for full details.

Another feature of these data is the presence of extreme observations, both at baseline and
Figure 1: Strip plots of IgG, IFNγ, IL-6, IL-10 and TNFα (in logarithmic scale) measured from PBMCs at baseline and after challenging with complexes of GRP94 and IgG. Values on the horizontal axis are (arbitrarily) ordered according to patient ID. Patient ID 5 was removed from the study and cytokines’ measurements for patients with ID 17, 27 and 28 are missing.

Such extreme observations induce high variability on the response measurements, especially for IFNγ, IL-6, IL-10 and TNFα. Hence, one must be cautious when fitting a LMM to such data; see Section 5 for the complete analysis of the GRP94 data.
3 Background

The present section provides an overview on concepts and methods that are needed to make the paper as much self-contained as possible. In particular, Section 3.1 gives background on robust $M$-estimation, Section 3.2 provides an overview on pseudo-likelihoods and Section 3.3 gives background on ABC.

3.1 Robust $M$-Estimation

Let $y = (y_1, \ldots, y_n)$ be a random sample of size $n$, having independent and identically distributed components, according to a distribution function $F_\theta = F(y; \theta)$, with $\theta \in \Theta \subseteq \mathbb{R}^d$, $d \geq 1$. Let $L(\theta; y)$ be the likelihood function based on model $F_\theta$.

Let $\Psi_\theta = \Psi(y; \theta) = \sum_{i=1}^n \psi(y_i; \theta)$ be an unbiased estimating function for $\theta$, i.e. such that $E_\theta(\psi(Y; \theta)) = 0$ for every $\theta$, with $\psi(\cdot)$ known function and with $E_\theta(\cdot)$ expectation with respect to $F_\theta$. Typically, $\Psi_\theta$ can be written as

$$
\Psi_\theta = b(\theta)^T a(y, \theta) - c(\theta),
$$

where $a(\cdot, \cdot)$ and $c(\cdot)$ are known vector-valued functions and $b(\theta)$ may be a known vector- or matrix-valued function of suitable dimensions. The function $c(\cdot)$ is typically a consistency correction, i.e. it ensures that $E_\theta(\Psi_\theta) = 0$. Most of the robust M-estimating functions can be recast in this form.

A general $M$-estimator (see, e.g., Hampel et al., 1986, Huber and Ronchetti, 2009) is defined as the root $\hat{\theta}$ of the estimating equation

$$
\Psi_\theta = 0.
$$

The class of $M$-estimators is wide and includes a variety of well-known estimators. For example, it includes the maximum likelihood estimator (MLE), the maximum composite likelihood estimator (see, e.g., Ruli et al., 2016, and references therein) and the scoring rule estimator (see Basu et al., 1998, Dawid et al., 2016, and references therein).

Under broad conditions assumed throughout this paper (see, e.g., Hampel et al., 1986, Huber and Ronchetti, 2009), an $M$-estimator is consistent and approximately normal with
mean $\theta$ and variance

$$K(\theta) = H(\theta)^{-1} J(\theta) H(\theta)^{-T},$$

where $H(\theta) = -E_\theta(\partial \Psi_\theta / \partial \theta^T)$ and $J(\theta) = E_\theta(\Psi_\theta \Psi_\theta^T)$ are the sensitivity and the variability matrices, respectively. The matrix $G(\theta) = K(\theta)^{-1}$ is known as the Godambe information and the form of $K(\theta)$ is due to the failure of the information identity since, in general, $H(\theta) \neq J(\theta)$.

The IF of the estimator $\tilde{\theta}$ is given by

$$IF(x; \tilde{\theta}, F_\theta) = H(\theta)^{-1} \psi(x; \theta),$$

and it measures the effect on the estimator $\tilde{\theta}$ of an infinitesimal contamination at the point $x$, standardised by the mass of the contamination. The supremum of the IF, i.e. the gross-error sensitivity, measures the worst influence on $\tilde{\theta}$ and a desirable robustness property for a statistical procedure is that the gross-error sensitivity is finite, i.e. that the IF is bounded (B-robustness). Note that the IF of the MLE is proportional to the score function; therefore, in general, MLE has unbounded IF, i.e. is not B-robust. On the contrary, if $\psi(x; \theta)$ is bounded, then the corresponding M-estimator $\tilde{\theta}$ is B-robust (Hampel et al., 1986). Finally, note that the IF can also be used to evaluate the asymptotic covariance matrix of $\tilde{\theta}$, since $K(\theta) = E_\theta(IF(x; \tilde{\theta}, F_\theta) IF(x; \tilde{\theta}, F_\theta)^T)$.

### 3.2 Pseudo-likelihoods from unbiased estimating functions

In general, there is not a unique function which has first derivative equal to $\Psi_\theta$, and many efforts have been made in order to derive pseudo-likelihood functions for $\theta$ based on $\Psi_\theta$. Pseudo-likelihoods are functions of the parameter $\theta$ and the data, with properties similar to a genuine likelihood function. Here we focus on three possible pseudo-likelihoods derived from $\Psi_\theta$.

A first pseudo-likelihood may be defined in terms of the conditional density of $\tilde{\theta}$ given $\theta$. In particular, assume that the joint density of $y$ is parametrized in terms of $\theta$ and a nuisance parameter $\lambda$ as

$$f(y; \theta, \lambda) = g(\tilde{\theta}; \theta) h(y|\tilde{\theta}; \theta, \lambda).$$
Then \( g(\hat{\theta}; \theta) \) is a pseudo-likelihood for \( \theta \) and its use is justified provided the information contained in \( h(y|\hat{\theta}; \theta, \lambda) \) about \( \theta \), when \( \lambda \) is unknown, is small or irrelevant (Davison et al., 1992). In practice \( g(\hat{\theta}; \theta) \) is unknown, but an approximation can be obtained via the bootstrap method (see, e.g., Davison et al., 1992, Boos and Monahan, 1986, Rubin, 1981), or by saddle-point approximations (see, e.g., Ronchetti and Welsh, 1994, and references therein). The restricted likelihood method of Lewis et al. (2014) can be seen as a simulation-based approximation of \( g(\tilde{\theta}; \theta) \).

Other pseudo-likelihoods, such as the quasi-likelihood (McCullagh, 1991, Adimari and Ventura, 2002b,a) and the empirical likelihood (Owen, 2001) provide other means of building pseudo-likelihood function directly from \( \Psi_\theta \).

The restricted likelihood method of Lewis et al. (2014) can be seen as a simulation-based approximation of \( g(\tilde{\theta}; \theta) \).

The quasi-likelihood is defined as

\[
L_Q(\theta) = \exp \left( \int_k^d A(t) \Psi(y; t) \, dt \right),
\]

where the matrix \( A(\theta) \) is such that \( A(\theta)^T = J(\theta)^{-1} H(\theta) \) and \( k \) is an arbitrary constant. When \( d = 1 \), \( L_Q(\theta) \) is usually easy to derive, but when \( d > 1 \) the integrals is path-dependent. The adjustment \( \Psi^a_\theta = A(\theta) \Psi(y; \theta) \) of \( \Psi_\theta \) is necessary in order to recover the information identity, and thus the correct curvature (see, e.g., Pace and Salvan, 1997, Chap. 4). Since \( |A(\theta)| \neq 0 \), \( \Psi^a_\theta = 0 \) leads to the same solution \( \tilde{\theta} \) of the original estimating equation \( \Psi_\theta = 0 \) and the robustness properties of \( \tilde{\theta} \) do not depend on \( A(\theta) \) because this matrix does not change its IF.

The empirical likelihood \( L_E(\theta) \) is defined through the empirical likelihood ratio statistic

\[
W_E(\theta) = -2 \log R(\theta), \quad 
R(\theta) = \max_{p_i} \prod_{i=1}^n n p_i,
\]

where the \( p_i \) weights satisfy \( p_i \geq 0 \), \( \sum_{i=1}^n p_i = 1 \) and \( \sum_{i=1}^n \psi(y_i; \theta)p_i = 0 \). A Lagrangian argument leads to

\[
W_E(\theta) = 2 \sum_{i=1}^n \log(1 + \eta^T \psi(y_i; \theta)),
\]

if \( \theta = 0 \) is inside the convex hull of \( \psi(y_1; \theta), \ldots, \psi(y_n; \theta) \); otherwise, it is adequate to set \( W_E(\theta) = +\infty \). The Lagrangian multiplier \( \eta = \eta(\theta) \) satisfies \( \sum_{i=1}^n \psi(y_i; \theta)/(1 + \eta^T \psi(y_i; \theta)) = 0 \).

Davison et al. (1992) show that, under suitable regularity conditions, the bootstrap likelihood is asymptotically equivalent to first-order to the empirical likelihood \( L_E(\theta) \). Moreover,
under standard regularity conditions, it can also be shown that the quasi-likelihood ratio statistic \( W_Q(\theta) \) and \( W_E(\theta) \) are equivalent to the first term of their Taylor expansions. Lastly, Monti and Ronchetti (1993) investigate connections between \( L_E(\theta) \) and saddlepoint pseudo-likelihoods in the case of \( M \)-estimating functions.

In the Bayesian framework, the use of \( L_Q(\theta) \) in place of the proper likelihood has been discussed by Greco et al. (2008) and Ventura et al. (2010), while the use of \( L_E(\theta) \) has been discussed by Lazar (2003), though not for robustness purposes, and by Greco et al. (2008). Since \( L_Q(\theta) \) and \( L_E(\theta) \), as well as the bootstrap and the saddle-point likelihoods, share most of the properties of the genuine likelihood, they can be used as a replacement of the latter in the Bayes’ formula (Ventura and Racugno, 2016) which, in the case of robust estimating functions, leads to the robust posterior distribution

\[
\pi_R(\theta|y) \propto \pi(\theta) L_R(\theta),
\]

where \( \pi(\theta) \) is a prior distribution for \( \theta \) and \( L_R(\theta) \) is a pseudo-likelihood based on a robust \( \Psi_\theta \), i.e. \( L_Q(\theta), L_E(\theta) \).

The approach based on robust posterior distributions (2) derived from robust \( M \)-estimating functions has two main drawbacks: the empirical likelihood is not computable for very small values of the sample size and for moderate sample sizes appears to have always heavy tails (see, e.g., Adimari and Ventura, 2002b, Greco et al., 2008); the posterior distribution based on the quasi-likelihood can be easily obtained only for scalar parameters. Lastly, efficient use of bootstrap, empirical saddle-point and of restricted likelihood methods requires \( M \)-estimating functions that are easy to evaluate at different parameter values, which is generally not the case, especially in the context of this paper.

3.3 Approximate Bayesian Computation

Given a prior \( \pi(\theta) \) and assuming that simulation from \( F_\theta \), at a given \( \theta \), is possible, the ABC method (Tavaré et al., 1997, Pritchard et al., 1999) can provide an approximation of the posterior distribution \( \pi(\theta|y) \), by means, for instance, of the accept-reject ABC algorithm (see Algorithm 1 and Tavaré et al., 1997). The latter samples from the joint distribution
for $i = 1 \rightarrow m$ do
repeat
\hspace{1em} draw $\theta^* \sim \pi(\cdot)$;
\hspace{1em} draw $y^* \sim F_{\theta^*}$
until $\rho(y, y^*) \leq \epsilon$;
set $\theta^{(i)} = \theta^*$
end

Algorithm 1: ABC accept-reject algorithm.

$$
\pi_\epsilon(\theta, y^*|y) = \frac{\pi(\theta)f(y^*; \theta)\mathbb{1}_{A_{\epsilon,y}(y^*)}}{\int_{A_{\epsilon,y} \times \Theta} \pi(\theta)f(y^*; \theta) dy^* d\theta},
$$

where $\mathbb{1}_{A_{\epsilon,y}(y^*)}$ is the indicator function of the set $A_{\epsilon,y}(y^*) = \{y^* : \rho(y^*, y) \leq \epsilon\}$, $\rho(\cdot)$ is a given distance and $\epsilon > 0$ is a fixed tolerance level. The ABC method provides an approximation of the posterior distribution $\pi(\theta|y)$, given by

$$
\pi_\epsilon(\theta|y) = \int \pi_\epsilon(\theta, y^*|y) dy^*.
$$

If $\epsilon \to 0$, then $\pi_\epsilon(\theta|y) \to \pi(\theta|y)$ (Blum, 2010). The threshold $\epsilon$ is generally set to the $\alpha$th quantile of the distance among the simulated and observed data, with $\alpha$ being typically very small (see, e.g., Beaumont et al., 2002).

In practice, the distance among the raw data can be rather noisy and a dimensional reduction of $y$ and $y^*$ is often necessary (see, e.g., Beaumont et al., 2002, Fearnhead and Prangle, 2012). This can be performed by considering a low-dimensional set of summary statistics $t(\cdot) = (t_1(\cdot), \ldots, t_q(\cdot))$, for some $q < n$, such as the mean, the median, quantiles etc. If $t(\cdot)$ is sufficient, then the ABC posterior $\pi_\epsilon(\theta|t(y))$ converges to the exact posterior as $\epsilon \to 0$ (see Soubeyrand et al., 2013, Appendix A). However, sufficient statistics are generally available only for models belonging to the exponential families.

The choice of the summary statistic is therefore a crucial point of ABC and this issue has been widely discussed in the statistical literature. In particular, Ruli et al. (2016, Sect. 3.1) show that, when a full computable likelihood is available, the standardised score function evaluated at the observed MLE $\hat{\theta}$, is an ideal summary statistic in ABC for a general model.
This ideal summary statistic is given by

$$\eta(y; \hat{\theta}) = B(\hat{\theta})^{-1} \ell_\theta(\hat{\theta}; y),$$

(3)

where $\ell_\theta(\theta; y) = \partial \log L(\theta; y)/\partial \theta$ is the score function and $B(\theta)$ is the square root of the expected information $i(\theta)$, i.e. is such that that $i(\theta) = B(\theta)B(\theta)^\top$. When considering the Mahalanobis distance we have

$$\rho(\eta(y; \hat{\theta}), \eta(y^*; \hat{\theta})) = \ell_\theta(\hat{\theta}; y^*)^\top i(\hat{\theta})^{-1} \ell_\theta(\hat{\theta}; y^*),$$

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

In situations in which the likelihood function is intractable, and a composite likelihood is available, Ruli et al. (2016, Sect. 3.2) show that the corresponding (suitably standardised) composite score function can be used as a summary statistic in ABC. The resulting ABC procedure is shown to be invariant to reparameterisations and it automatically adjusts the curvature of the composite likelihood and of the corresponding posterior distribution. Lastly, note that in practice the ABC accept-reject algorithm is computationally inefficient and more advanced algorithms, such as ABC coupled with MCMC (ABC-MCMC) (Marjoram et al., 2003), are preferred.

4 Robust ABC Bayesian inference

4.1 Methodology

One possibility to perform robust Bayesian inference is to resort to a pseudo-posterior distribution of the form (2). However, as discussed in Section 3.2, this approach has several limitations. A further computational limitation is that robust estimating equations $\Psi_\theta$ typically involve consistency corrections which are often intractable functions of $\theta$, defined as suitable expectations under the assumed model. See Section 4.4 for a simple illustrative example.

Here we propose an alternative method for computing posterior distributions based on robust $M$-estimating functions. The method resorts to the ABC machinery in which a
standardised version of $\Psi$ is used as summary statistic. The idea is an extension of Ruli et al. (2016) to the context of general unbiased estimating functions, though from a different perspective. The rescaled $M$-estimating function is

$$
\tilde{\eta}_R(y^*) = \eta_R(y^*; \tilde{\theta}) = B_R(\tilde{\theta})^{-1} \Psi(y^*; \tilde{\theta}),
$$

where $B_R(\theta)$ is such that $J(\theta) = B_R(\theta)B_R(\theta)^T$ and $\tilde{\theta} = \tilde{\theta}(y)$ is based on the observed sample $y$. To generate posterior samples we propose to use the ABC algorithm with an MCMC kernel (see Algorithm 2), similar to Algorithm 2 of Fearnhead and Prangle (2012); see also Marjoram et al. (2003). In Algorithm 2, $K_h(\cdot)$ denotes a density kernel with bandwidth $h$.

The latter controls discrepancy between the observed and simulated summary statistics; in particular, the lower is $h$ the lower is the discrepancy.

**Result:** A Markov dependent sample $(\theta^{(1)}, \ldots, \theta^{(m)})$ from $\pi^R_{\text{ABC}}(\theta|\tilde{\theta})$.

**Data:** a starting value $\theta^{(0)}$, a proposal density $q(\cdot|\cdot)$

**for** $i = 1 \rightarrow m$ **do**

- draw $\theta^* \sim q(\cdot|\theta^{(i-1)})$
- draw $y^* \sim F_{\theta^*}$
- draw $u \sim U(0, 1)$

  **if** $u \leq \frac{K_h(\tilde{\eta}_R(y^*))}{K_h(\tilde{\eta}_R(y^{(i-1)}))} \frac{\pi(\theta^*)q(\theta^{(i-1)}|\theta^*)}{\pi(\theta^{(i-1)})q(\theta^*|\theta^{(i-1)})}$ **then**
  - set $(\theta^{(i)}, \tilde{\eta}_R^{(i)}) = (\theta^*, \tilde{\eta}_R(y^*))$
  **else**
  - set $(\theta^{(i)}, \tilde{\eta}_R^{(i)}) = (\theta^{(i-1)}, \tilde{\eta}_R(y^{(i-1)}))$

**end**

**Algorithm 2:** ABC-R algorithm with MCMC.

The proposed method via Algorithm 2 gives Markov-dependent samples from the ABC-R posterior

$$
\pi^R_{\text{ABC}}(\theta|\tilde{\theta}) = \frac{\int_{Y^*} \pi(\theta) f(y^*; \theta) K_h(\tilde{\eta}_R(y^*)) \, dy^*}{\int_{Y^* \times \Theta} \pi(\theta) f(y^*; \theta) K_h(\tilde{\eta}_R(y^*)) \, dy^* \, d\theta}.
$$

Using similar arguments to Soubeyrand et al. (2013, Appendix A), it can be shown that, for $h \rightarrow 0$, $\pi^R_{\text{ABC}}(\theta|\tilde{\theta})$ converges to $\pi(\theta|\tilde{\theta})$ pointwise (see also Blum, 2010), in the sense
that $\pi_{ABC}^{R}(\theta|\tilde{\theta})$ and $\pi(\theta|\tilde{\theta})$ are equivalent for sufficiently small $h$. Since in general (4) is not sufficient, then $\pi(\theta|\tilde{\theta})$ may differ from $\pi(\theta|y)$ and information is lost by using (4) instead of $y$. However such a loss of efficiency pays off in terms of robustness in the inference about $\theta$.

Posterior conditional on partial information have been extensively discussed in the literature (see, e.g., Doksum and Lo, 1990, and reference therein). Soubeyrand and Haon-Lasportes (2015) study the properties of the ABC posterior when the summary statistic is the MLE or the pseudo-MLE derived from a simplified parametric model. However, they do not treat the case with the summary statistic given by an estimating function. Indeed, an alternative version of the ABC-R algorithm could be based on $\tilde{\theta}$ used as the summary statistic and a, possibly rescaled, distance among the observed and the simulated statistic as in Soubeyrand and Haon-Lasportes (2015). Apparently, these two versions of ABC, namely the one based on $\tilde{\theta}$ and that based on (4) seem to be treated in the literature as two separate approaches (see, e.g., Gleim and Pigrosch, 2013, Drovandi et al., 2015). However, both methods use essentially the same statistic $\tilde{\theta}$, but through different distance metrics. In addition, for small tolerances $\epsilon$ and $h$, these two distances converge to zero, and both methods give a posterior distribution conditional on the same statistic $\tilde{\theta}$. Indeed, let $\tilde{\theta}$ be the summary statistic of the usual ABC posterior and let the tolerance threshold $\epsilon$ be sufficiently small and consider the random draw $\theta^*$ and its corresponding simulated summary statistics $\tilde{\theta}^*$ taken with the ABC algorithm. Then, by construction $\tilde{\theta}^*$ will be close to $\tilde{\theta}$. This implies that also $\eta_R(y^*; \tilde{\theta})$ will be close to $\eta_R(y^*; \tilde{\theta}^*) = 0$, and hence $\theta^*$ is also a sample from the ABC-R posterior which uses the summary statistic $\eta_R$.

The use of $\tilde{\theta}$ as summary statistic in ABC requires the solution of $\Psi_{\theta} = 0$ at each iteration, which is clearly computationally cumbersome. Therefore, this alternative approach will not be pursued further. Moreover, the proposed approach, besides sharing the same invariance properties stated by Ruli et al. (2016), i.e. invariance with respect to both monotonic transformation of the data or with respect to reparametrisations, has the advantage of avoiding some computational problems related to the evaluation of $\Psi_{\theta}$ as shown by the following lemma.

**Lemma 4.1** The ABC-R algorithm does not require repeated evaluations of the consistency
correction involved \( c(\theta) \) in \( \Psi_\theta \) as given by (3.1).

**Proof.** Let \( \tilde{\theta} \) be the solution of \( \Psi_{\theta} = 0 \), with \( \Psi_{\theta} \) of the form (3.1). Then, for a given simulated \( y^* \) at \( \theta^* \), we have

\[
\tilde{\eta}_R(y^*) = B_R(\tilde{\theta})^{-1}(\Psi(y; \tilde{\theta}) - \Psi(y^*; \tilde{\theta})) = b(\tilde{\theta})^\top \{a(y, \tilde{\theta}) - a(y^*, \tilde{\theta})\}.
\]

This implies that \( c(\theta) \), as also \( b(\theta) \), is computed only once, at \( \tilde{\theta} \).

### 4.2 Theoretical properties

Theorem 4.2 below shows that the proposed method gives a robust approximate posterior distribution with the correct curvature, even though \( \Psi_{\theta} \), unlike the full score function, does not satisfy the information identity.

**Theorem 4.2** The ABC-R algorithm with rescaled M-estimating function \( \eta_R(y) \), as \( h \to 0 \), leads to an approximate posterior distribution with the correct curvature and is also invariant to reparameterisations.

**Proof** The proof follows from Theorem 3.2 of Ruli et al. (2016), by substituting the composite estimating equation with the more general M-estimating function \( \Psi_{\theta} \).

The ABC-R algorithm delivers thus a robust approximate posterior distribution which does not need calibration. This is not the case with (2) for which a calibration is typically required.

The following Theorem 4.3 shows that the proposed ABC approximate posterior distribution is asymptotically normal. In the theorem, the regularity assumptions of Soubeyrand and Haon-Lasportes (2015) are assumed, as well as the Euclidean distance in the ABC-R algorithm. Under these assumptions the density of \( \tilde{\theta} \) is, for large \( n \), equivalent to the normal density with mean vector equal to \( \theta \) and covariance matrix equal to \( K(\theta) \), and the density of \( \Psi_{\theta} \) is, for large \( n \), equivalent to the normal density with zero mean vector and covariance matrix equal to \( J(\theta) \).
**Theorem 4.3** Consider the ABC-R algorithm that samples from the posterior \( \pi_R^{ABC}(\theta|\tilde{\theta}) \). Assume that when \( h \to 0 \), \( \pi_R^{ABC}(\theta|\tilde{\theta}) \) converges pointwise to \( \pi_R(\theta|\tilde{\theta}) \). Moreover, when \( n \to \infty \) and \( h \to 0 \), the posterior \( \pi_R^{ABC}(\theta|\tilde{\theta}) \) is asymptotically equivalent to the density of the normal distribution with mean vector \( \tilde{\theta} \) and covariance matrix \( K(\tilde{\theta}) \):

\[
\pi_R^{ABC}(\theta|\tilde{\theta}) \sim N_d(\tilde{\theta}, K(\tilde{\theta})), \quad \text{for } h \to 0 .
\]

**Proof.** The proof follows from Lemma 2 and Theorem 1 in Soubeyrand and Haon-Lasportes (2015) and from the asymptotic relation between the Wald-type statistic and the score-type statistic, i.e.

\[
\tilde{\eta}_R(y)^T \tilde{\eta}_R(y) = \Psi_\theta^T J(\theta)^{-1} \Psi_\theta = (\tilde{\theta} - \theta)^T K(\tilde{\theta})^{-1}(\tilde{\theta} - \theta) + o_p(1) .
\]

In view of Theorem 4.3, the resulting point estimates and credible sets converge to their frequentist analogues.

If \( \psi(y; \theta) \) is bounded in \( y \), i.e. if the estimator \( \tilde{\theta} \) is B-robust, then the ABC-R posterior is resistant with respect to slight violations of the model assumptions. The advantage of the ABC-R posterior with respect to a robust posterior distribution of the form (2) is that the former allows to easily deal with multidimensional parameters, and thus avoids the difficulties of the quasi-likelihood. Moreover, it is computable also for very small sample sizes, provided \( \tilde{\theta} \) is finite, and thus avoids the numerical instabilities of the empirical likelihood.

**Theorem 4.4** If \( \psi(y; \theta) \) is bounded in \( y \), i.e. if the estimator \( \tilde{\theta} \) is B-robust, then asymptotically the posterior mode, as well as posterior moments, of \( \pi_R^{ABC}(\theta|\tilde{\theta}) \) have bounded IF.

**Proof.** From Theorem 4.3, the asymptotic posterior mode of \( \pi_R^{ABC}(\theta|\tilde{\theta}) \) is \( \tilde{\theta} \), which is B-robust. Moreover, following results in Greco et al. (2008), it can be shown that asymptotic posterior moments have bounded IF if and only if the posterior mode has bounded IF.

Notice that the ABC method can be seen as inherently robust since the resulting posterior conditions on a user-specified subset of the data instead of on the full data, as in the classical Bayesian setting. Nevertheless, through the ABC-R method we are able to obtain a B-robust posterior, that is a posterior having summaries with a bounded influence function.
4.3 Implementation details

Provided simulation from $F_{\theta}$ is fast, the main demanding requirement of the proposed method is essentially the computation of the observed $\tilde{\theta}$ and the scaling matrix $B_{R}(\theta)$ evaluated at $\tilde{\theta}$.

The ABC-R Algorithm 2 involves a kernel density $K_{h}(\cdot)$, which is governed by the bandwidth $h$ and a proposal density $q(\cdot|\cdot)$. Given that, for large sample sizes,

$$\tilde{\eta}_{R}(y) \sim N(d, I_{d})$$

where $0_{d}$ is a $d$-vector of zeros and $I_{d}$ is the identity matrix of order $d$, it is reasonable to replace $K_{h}(\cdot)$ with the multivariate normal density centred at zero and with covariance matrix $hI_{d}$. To choose the bandwidth $h$ we consider several pilot runs of the ABC-R algorithm for a grid of $h$ values, and select the value of $h$ that delivers approximately 1% acceptance ratio (as done, for instance, by Fearnhead and Prangle, 2012).

Contrary to other ABC-MCMC settings in which the proposal requires pilot runs (see, Cabras et al., 2015, for building proposal distributions in ABC), in our case a scaling matrix for the proposal $q(\cdot|\cdot)$ can be readily build, almost effortlessly, by using the usual sandwich formula (1) (see also Ruli et al., 2016) evaluated at $\tilde{\theta}$. Even in cases in which $H(\theta)$ and $J(\theta)$ are not analytically available, they can be straightforwardly estimated via simulation. Indeed, in our experience, 100-500 samples from the model $F_{\theta}$, with $\theta = \tilde{\theta}$, give estimates with reasonably low Monte Carlo variability (see also Cattelan and Sartori, 2015). Throughout the examples considered we use the multivariate $t$-density with 5 degrees of freedom as the proposal density $q(\cdot|\cdot)$ and the ABC-R is always started from $\tilde{\theta}$. An R implementation of the proposed method for linear mixed effects models is also provided through the robustBLME package available on CRAN.

As a final remark, we note that robust estimating functions are available for many models of practical interest; see, among others, Rousseeuw and Leroy (2003), Heritier et al. (2011), Huber and Ronchetti (2009), Farcomeni and Ventura (2012) and Farcomeni and Greco (2015) and references therein. Therefore, it is straightforward to extent the application of the proposed method in such model settings.
4.4 An illustrative example

We consider an example in which we compare the numerical properties of the ABC-R posterior, with the classical posterior based on the assumed model and the pseudo-posterior based on the empirical likelihood. Scenarios with data simulated either from the assumed model or from a slightly misspecified model are considered.

Let $F_{\theta}$ be a location-scale distribution with location $\mu$ and scale $\sigma > 0$, and let $\theta = (\mu, \sigma)$. $M$-estimation through Huber’s estimating function is a standard choice for the robust estimation of location and scale parameters. The $M$-estimating function is

$$\Psi_\theta = (\Psi_\mu, \Psi_\sigma),$$

with

$$\Psi_\mu = \sum_{i=1}^{n} \psi_{c_1}(z_i) \quad \text{and} \quad \Psi_\sigma = \sum_{i=1}^{n} \left( \psi_{c_2}(z_i)^2 - k(c_2) \right),$$

(7)

where $z_i = (y_i - \mu)/\sigma$, $i = 1, \ldots, n$, $\psi_{c_1}(z) = \min(1; c_1/|z|)$ is the Huber $\psi$-function, $c_1 > 0$ and $c_2 > 0$ are scalar parameters which control the desired degree of robustness of $\hat{\theta}$, and $k(\cdot)$ is a known function.

As an example let $F_\theta$ be the normal distribution $N(\mu, \sigma^2)$ and assume $\mu$ and $\sigma$ a priori independent with $\mu \sim N(0, 10^2)$ and $\sigma \sim \text{halfCauchy}(5)$, where halfCauchy($a$) is the half Cauchy distribution with scale parameter equal to $a$. We consider random samples of sizes $n = \{15, 30\}$ drawn from either the normal distribution with $\theta = (0, 1)$ and from a contaminated model. The latter is given by $F_\epsilon = (1 - \epsilon)N(0, 1) + \epsilon N(0, \sigma_1^2)$, with $\sigma_1^2 > 0$. For illustration purposes, we set the contamination to 10%, i.e. $\epsilon = 0.1$, and we set $\sigma_1^2 = 10$. Moreover, we assume $c_1 = 1.345$ and $c_2 = 2.07$, which imply that $\tilde{\mu}$ and $\tilde{\sigma}$ are, respectively, 5% and 10% less efficient than the corresponding MLE under the central (normal) model.

In the simulation study we compute the genuine posterior distribution, e.g. the posterior based on the likelihood function of the normal model, the pseudo-posterior (2) based on the empirical likelihood (EL) and the ABC-R posterior.

The genuine and the EL posteriors are computed by numerical integration. The ABC-R posterior is obtained using Algorithm 2 (Section 4.3). From the posterior distributions
illustrated in Figure 2 we note that, when the data come from the central model (panels (a)-(b)), i.e. for $\epsilon = 0$, all the posteriors are in reasonable agreement, even if the EL posterior behaves slightly worse, especially the marginal posterior of $\sigma$ with $n = 15$. When the data are contaminated (panels (c)-(d)), the genuine posterior is less trustworthy as the bulk of the posterior drifts away from the true parameter value (vertical and horizontal straight lines). This is not the case however for the ABC-R posterior which remains centered around the true parameter value. The behavior of the EL posterior is less clear-cut here. In particular, it seems to have bimodal shape with the dominant mode laying around the true value and the secondary mode being located distantly. Consequently, the EL posterior has higher spread than the ABC-R posterior.

**Sensitivity analysis**

To assess the robustness of the ABC-R posterior, we consider a sensitivity analysis. We simulate a sample $y$ of size $n = 31$ from the central model and the contaminated data are given by $y^c = (y_{(j)}, y_{(n/2)} + c)$, $j = 1, \ldots, n/2 - 1, n/2 + 1, \ldots, n$, where $y_{(j)}$ denotes the $j$th ordered value of $y$ and $c$ a contamination scalar with possible values $\{-15, -14, \ldots, 15\}$. The results of the sensitivity analysis, illustrated by means of violin plots in Figure 3, highlight that the posterior median (the symbol ”◦”) of the genuine posterior (panel (c)) is substantially driven by $c$. The $y$-axis is left free in order to enhance the readability of the plot. On the other hand, the ABC-R and the EL posteriors are more resistant. For all the posteriors, the behavior of the posterior median reflects the behavior of the influence function of the posterior mode. Finally, the EL posterior under contamination is much more dispersed than the ABC-R posterior. Hence, even in an example as simple as this, robust inference based on the ABC-R posterior appears to be more precise than inference based on the EL posterior.
Figure 2: First row: genuine (continued), EL (blue dashed) and ABC-R posteriors (shaded image and histogram) for the normal model, when the data come from the central model $N(0,1)$ with (a) $n = 15$ and (b) $n = 30$. Second row: genuine, EL and ABC-R posteriors for the normal model, when the data come from the contaminated model with $\epsilon = 0.1$, (c) $n = 15$ and (d) $n = 30$. 
Figure 3: Marginal ABC-R (a), EL (b) and genuine (c) posteriors for $\mu$ (left columns) and $\sigma$ (right) represented by means of violin plots for each $y^c$ generated under each value of $c = \{-15, -14, \ldots, 15\}$. For each violin plot, the central circle represents the posterior median. The horizontal lines are the corresponding posterior median under $y^c$ with $c = 0$. 
5 Application to LMM

5.1 Classical and Robust Estimation of LMM

The linear mixed model is a popular framework for analysing datasets in the context of hierarchical, longitudinal or repeated measures. A general formulation of LMM is by the following regression equation

\[ y = X\alpha + \sum_{i=1}^{c-1} Z_i\beta_i + \varepsilon, \]  

where \( y \) is a \( n \)-vector of the response observations; \( X \) and \( Z_i \) are known \( n \times q \) and \( n \times p_i \) design matrices respectively; \( \alpha \) is a \( q \)-vector of unknown fixed effects; the \( \beta_i \) are \( p_i \)-vectors of unobserved random effects \( (1 \leq i \leq c - 1) \); and \( \varepsilon \) is an \( n \)-vector of unobserved errors.

The \( p_i \) levels of each random effect \( \beta_i \) are assumed to be independent with mean zero and variance \( \sigma_i^2 \). Moreover, each random error \( \varepsilon_i \) is assumed to be independent with mean zero and variance \( \sigma_c^2 \) and \( \beta_1, \ldots, \beta_{c-1} \) and \( \varepsilon \) are assumed to be independent.

In this paper we focus on the classical normal LMM, which assumes that \( \varepsilon \sim N_n(0_n, \sigma_c^2 I_n) \) and \( \beta_i \sim N(0, \sigma_i^2) \), \( i = 1, \ldots, c - 1 \). Obviously, the validity and performance of this LMM requires strict adherence to the model assumption, which is usually chosen because it simplifies the analyses and not because it fits exactly the data at hand. The robust procedure discussed in this paper specifically takes into account the fact that the normal model is only approximate and then it produces statistical analyses that are stable with respect to outliers, deviations from the model or model misspecifications.

Under (8) and the aforementioned assumptions, it follows that \( Y \) is multivariate normal with

\[ E(Y) = X\alpha \quad \text{and} \quad \text{var}(Y) = V = \sum_{i=1}^{c-1} \sigma_i^2 Z_i Z_i^T, \]

where \( Z_c = I_n \). We assume that the set of \( d = q + c \) unknown parameters \( \theta = (\alpha, \sigma^2) = (\alpha, \sigma_1^2, \ldots, \sigma_c^2) \) is identifiable.

Although the \( n \) observations \( y \) are not independent, if the random effects are nested, then independent subgroups of observations can be found. Indeed, in many situations, \( y \) can be split into \( g \) independent groups of observations \( y_j, j = 1, \ldots, g \). In this case the log-likelihood function is
\[
\ell(\theta) = (-1/2) \sum_{j=1}^{g} \left\{ \log|V_j| + (y_j - X_j \alpha)^T V_j^{-1} (y_j - X_j \alpha) \right\},
\] (9)

where \((y_1, \ldots, y_g), X\) and \(V\) are partitioned conformably.

Classical Bayesian inference for \(\theta\) is based on the posterior distribution \(\pi(\theta|y) \propto \exp\{\ell(\theta)\} \pi(\theta)\), where \(\pi(\theta)\) is a prior distribution for \(\theta\). However, \(\ell(\theta)\) in the Gaussian LMM (9) can be very sensitive to model deviations. This is because in the LMM framework a strict adherence to the multivariate normal model is required, and even a mild deviation from this model can have a great impact on the inferential results (Richardson and Welsh, 1995, Richardson, 1997, Copt and Victoria-Feser, 2006); see also results of the simulation study in Section 5.2.

There is a rich variety of robust inferential procedures for LMM in the presence of model misspecification: \(M\)-estimators (see, e.g., Richardson and Welsh, 1995, Richardson, 1997, Welsh and Richardson, 1997, and references therein), \(S\)-estimators (Copt and Victoria-Feser, 2006), \(MM\)-estimators (Copt and Heritier, 2006) and multivariate \(t\) distributions (see, e.g., Lange et al., 1989). The \(M\)-, \(S\)- and \(MM\)-estimation in the LMM framework provide robust estimators derived from estimating equations which are unbiased under the central model. In particular, \(S\)- and \(MM\)-estimators provide robust estimators with high breakdown point, but they are generally available only for balanced designs. Robustness and efficiency in \(M\)-, \(S\)- and \(MM\)-type estimators is guided by tuning constants which in practice are fixed by choosing the desired amount of robustness or, equivalently, by setting the desired loss of efficiency with respect to the MLE (see also Basu et al., 1998). The multivariate \(t\) framework deals with robustness by replacing the normal model with the multivariate \(t\), but at the cost of introducing the degrees of freedom as an additional unknown parameter.

Here we focus on robust inference in LMM through the \(M\)-estimation approach as pioneered by Richardson and Welsh (1995), Richardson (1997) and Welsh and Richardson (1997) and extend it to the Bayesian framework through the ABC-R method. The aforementioned \(M\)-estimation approach is very general as it can deal with a wide variety of situations; for instance it can deal with unbalanced designs and robustness with respect to the design matrix (see Richardson, 1997, for further details).

Motivated by the GRP94 dataset of Section 2, in the present work we focus on robust-
ness with respect to the response variable. Following Welsh and Richardson (1997) (and Richardson, 1997), we focus on bounded of M-estimating functions of the form

\[ X^T V^{-1/2} \psi_{c_1} \left( V^{-1/2} (y - X\alpha) \right) = 0, \]  

(10a)

\[ (1/2) \left\{ \psi_{c_2} \left( V^{-1/2} (y - X\alpha) \right)^T V^{-1/2} Z_i Z_i^T V^{-1/2} \right. \]

\[ \left. \times \psi_{c_2} \left( V^{-1/2} (y - X\alpha) \right) - \text{tr}(CPZ_i Z_i^T) \right\} = 0, \text{ for } i = 1, \ldots, c \]

(10b)

where \( C = E(\psi_{c_2}(R)\psi_{c_2}(R)^T) \), with \( R = V^{-1/2}(Y - X\alpha) \), \( P = V^{-1} - V^{-1}X(X^TV^{-1}X)^{-1}X^TV^{-1} \) and \( \text{tr}(\cdot) \) is the trace operator. The function \( \text{tr}(CPZ_i Z_i) \) is a correction factor needed to ensure consistency at the Gaussian model for each \( i = 1, \ldots, c \).

Welsh and Richardson (1997) call (10a)-(10b) robust REML II estimating equations as they are bounded versions of restricted likelihood equations (Patterson and Thompson, 1971). Richardson (1997) shows that the estimator of \( \theta \) given by the solution of (10a)-(10b) is asymptotically normal with mean the true parameter \( \theta \) and variance matrix of the form (1).

### 5.2 Simulation study

Let us consider the following two-component nested model

\[ y_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij}, \]  

(11)

where \( \mu \) is the grand mean; \( \alpha_j \) are the fixed effects, constrained such that \( \sum_{j=1}^q \alpha_j = 0 \); \( \beta_i \sim N(0, \sigma_1^2) \) are the random effects and \( \varepsilon_{ij} \sim N(0, \sigma_2^2) \) is the residual term, for \( j = 1, \ldots, q \) and \( i = 1, \ldots, g \). This model is a particular case of (8), with \( c = 2 \), a single random effect \( \beta = \beta_1 \) with \( p_1 = g \) levels and \( Z_1 \) being the unit diagonal matrix. Furthermore, the covariate is a categorical variable with \( q \) levels hence the design matrix is given by \( q - 1 \) dummy variables.

We assess bias and efficiency of the proposed method via simulations with 500 Monte Carlo replications. For each Monte Carlo replication, the true values for the variance components \((\sigma_1^2, \sigma_2^2)\) and those of \( \alpha \) are drawn uniformly in \((1, 10) \times (1, 10) \) and \((-5, 5)\), respectively. With these parameter values, two datasets each of size \( g \) are generated: one from the central
model and one from the contaminated model $F_{\epsilon} = (1 - \epsilon)N(X_i^\top \alpha, V_i) + \epsilon N(X_i^\top \alpha, 15V_i)$, with $\epsilon = 0.10$. We consider $q = \{3, 5, 7\}$ and $g = \{30, 50, 70\}$. Prior distributions are $\alpha \sim N_q(0, 10^2 I_q)$ and $(\sigma_1^2, \sigma_2^2) \sim \text{halfCauchy}(7) \times \text{halfCauchy}(7)$. For each scenario, we fit model (11) in the classical Bayesian way, using an adaptive random walk Metropolis-Hastings algorithm. The same model is fitted by the ABC-R method using the estimating equations (10a)-(10b). As in Richardson and Welsh (1995), we set $c_1 = 1.345$ and $c_2 = 2.07$. To find $\tilde{\theta}$, we solve (10a)-(10b) iteratively until convergence, as suggested by Richardson and Welsh (1995). The classical REML estimate, computed by the lmer function of the lme4 package (Bates et al., 2015), is used as starting value. In our experiments, the convergence of the solution is quite rapid, i.e. $\tilde{\theta}$ stabilises within 10 to 15 iterations.

We assess the component-wise bias of the posterior median $\tilde{\theta}_m$ by the modulus of $\tilde{\theta}_m - \theta_0$ in logarithmic scale, where $\theta_0$ is the true value. Moreover, the efficiency of the classical Bayesian estimator relative to the ABC-R estimator is assessed through the index $MD_{MCMC}/MD_{ABC}$, where $MD = \text{med}(\tilde{\theta}_m - \theta_0)$. A similar index is used also by Richardson and Welsh (1995) (see also Copt and Victoria-Feser, 2006). In addition, for each Monte Carlo replication we compute the Euclidean distance of $\tilde{\theta}_m$ from $\theta_0$, which can be considered as a global measure of bias. Contrary to Richardson and Welsh (1995), we consider a different $\theta_0$ for each Monte Carlo replication. For this reason, instead of just summarising the distribution of bias and efficiency by simple statistics we find it more informative to depict them graphically. The bias and efficiency of the classical Bayesian posterior and of the ABC-R posterior for the 500 replications are illustrated in Figures 4 and 5, respectively.

Under the central model, inference with the ABC-R and the classical Bayesian posteriors is roughly similar, i.e. both bias (component-wise and global) and efficiency compare equally well across the two methods. This holds for both the fixed effects $\alpha$ and for the variance components $\sigma_1^2, \sigma_2^2$. The latter present more bias than the former. Under the contaminated model, we notice important differences among ABC-R and the classical Bayesian estimation. In particular, $\tilde{\theta}_m$ based on ABC-R is less biased, both globally and on a component by component basis, and more efficient. The gain in efficiency is particularly evident for the variance components.
5.3 Effects of GRP94-based complexes on IgG and cytokines

We fit the two-component nested LMM (11) with ABC-R using estimating equations (10a)-(10b) to each of the five response variables: IgG, IFNγ, IL-6, IL-10 and TNFα. Since all measures are positive and some of them are highly skewed, a logarithmic transformation is used in order to alleviate distributional skewness. Furthermore, since Tramentozzi et al. (2016) highlight a possible gender effect (especially with respect to the cytokines) we check for gender effects by including an interaction with gender. In the multivariate form, the model with interaction can be written as

\[ y_i = X_i^T \alpha + X_i^T \times w_i \gamma + \beta_i 1_6 + \varepsilon_i, \quad i = 1, \ldots, g \]  \hspace{1cm} (12)

where \( X_i \) is the matrix of covariates for the \( i \)th unit, \( w_i \) is a dummy variable for gender, \( \gamma \) is the fixed effect of the treatment-gender interaction, \( 1_6 \) is the unit vector of dimension 6. The number of statistical units \( g \) is equal to 27 for the response IgG and 24 for the other response variables. Notice that the interaction model (12) has 12 unknown fixed effects \( (\alpha, \gamma) \).

As there is no extra-experimental information, we use relatively diffused priors. In particular, we assume \( \alpha_j \sim N(0, 100) \) and \( \gamma_j \sim N(0, 100) \), for \( j = 1, \ldots, 6 \). For the variance components, following the recommendation of Gelman (2006), we assume \( \sigma_1^2 \sim \text{halfCauchy}(7) \) and \( \sigma_2^2 \sim \text{halfCauchy}(7) \) in both models.

ABC-R posterior samples are drawn using Algorithm 2 and following the implementation details of Section 4.3. For comparison purposes, we fit also a classical Bayesian LMM with the aforementioned prior; an adaptive random walk Metropolis-Hastings algorithm is used for sampling from this posterior. Figure 6 compares the ABC-R and the classical posterior for a subset of the fixed effects of models (11) and (12) by means of kernel density estimations. The parameters shown are those referring to the treatments based on GRP94 at the dose of 10 ng/ml (GRP94_{10}), GRP94 at the dose of 100 ng/ml (GRP94_{100}) and GRP94 in complex with IgG at the dose of 100 ng/ml (GRP94+IgG_{100}), which according to Tramentozzi et al. (2016) are the most prominent. In Figure 6, panels (a1) to (e1) report the marginal posteriors of the fixed effects of (11) (with baseline being the reference category), fitted to the five response variables, respectively; panels (a2) to (e2) give those of (12) (with baseline and female being the reference categories). Numbers within parenthesis in the plot sub titles give
the evidence in favour of the null hypothesis $H_0$ that the parameter is equal to zero, computed under the Full Bayesian Significance Testing (FBST) setting of Pereira et al. (2008); inside the parenthesis, the first (last) value from left refers to the ABC-R (classical) posterior. The FBST in favour of $H_0$ has been proposed by (Pereira and Stern, 1999) as an intuitive measure of evidence, defined as the posterior probability related to the less probable points of the parametric space. It favours the null hypothesis whenever it is large and it is based on a specific loss function and thus the decision made under this procedure is the action that minimises the corresponding posterior risk (see Pereira et al., 2008, and references therein). The FBST solves the drawback of the usual Bayesian procedure for testing based on the Bayes factor (BF), that is, when the null hypothesis is precise and improper or vague priors are assumed, the BF can be undetermined and it can lead to the so-called Jeffreys-Lindleys paradox.

In panel (a1), the first plot from left reports the marginal distributions of the expected difference in terms of IgG between treatment $GRP94_{10}$ and baseline. It shows that such a treatment is quite different from baseline measurements, because under the classical Bayesian LMM the evidence is zero to three decimal places, whereas under the ABC-R posterior the evidence is as low as 0.002. In what follows, evidences greater than 0.15 are treated as not worth mentioning and are therefore omitted. From the plots we can conclude the following.
Figure 4: Bias of the ABC-R and classical (MCMC) Bayesian estimation of LMM under either the central (Full) or the contaminated model (Mix) for varying $n$ and $q$. Rows refer to a parameter or combination of parameters (row all_par); columns within each cell refer to different values of $q$; e.g. the last two rows (starting from top) have only two boxplots since $\alpha_6$ and $\alpha_7$ are available only with $q = 7$. 

27
Figure 5: Efficiency of the ABC-R compared to the classical Bayesian estimation of LMM under the central (Full) and the contaminated models (Mix) for varying $n$ and $q$. Rows refer to a parameter and columns within each cell refer to different values of $q$; e.g. the last two rows (starting from top) have only two boxplots since $\alpha_6$ and $\alpha_7$ are available only with $q = 7$. 

28
(a1) log IgG (marginal effects)

(a2) log IgG (interaction effects)

(b1) log (1+IFNg) (marginal effects)

(b2) log (1+IFNg) (interaction effects)

(c1) log IL–6 (marginal effects)

(c2) log IL–6 (interaction effects)
Figure 6: Comparison of robust (ABC) and full (MCMC) posterior distributions of the fixed effects of the LMM \((11)-(12)\), fitted to: log IgG (panels (a1)-(a2)), log(INF\(\gamma+1\)) ((b1)-(b2)), log IL-6 ((c1)-(c2)), log IL-10 ((d1)-(d2)) and log TNF\(\alpha\) ((e1)-(e2)). At each row, the left block refers to the posterior of the marginal effects of the treatments against the baseline and the right block refers to the interactions of the treatments with gender (with baseline and female being the reference categories). Numbers within refer to the evidence in favour of the null hypothesis \(H_0\) that the parameter is equal to zero, computed under the Full Bayesian Significance Testing (FBST) setting of Pereira et al. (2008); inside the parenthesis, the first (last) value from left refers to the ABC-R (classical) posterior. Dashed vertical lines corresponds to \(\tilde{\theta}\).
**IgG.** There is a high posterior probability that the marginal treatments are different from the baseline, since the evidence of $H_0$ is rather low. However, the interaction with gender seem not to be supported. Although the density plots reveal some qualitative differences between the classical and the proposed robust Bayesian LMM, the evidences in favour of $H_0$ according to these posteriors, are rather low and in reasonable agreement, hence the message they convey is the same.

**IFN$\gamma$.** No marginal nor interaction effects seem to be supported by the data for this variable according to both the classical and the proposed robust method.

**IL-6.** The classical Bayesian LMM suggests that there is a marginal effect of GRP94\_10 and of GRP94\_100 as in both cases the evidence is low, i.e. less than 0.05. Furthermore, there is also evidence of an interaction effect of the latter with gender. However, on the basis of the ABC-R posterior, only GRP94\_100 shows a marginal effect with respect to the baseline and no interaction effects are worth mentioning. These apparently contradictory results can be explained by the presence of few patients with extreme observations (see Figure 1) which seem to drive the classical LMM posterior but not the ABC-R posterior.

**IL-10 and TNF$\alpha$.** For both these variables we notice both marginal and interaction effects of GRP94\_100 under the classical Bayesian LMM. However, such effects disappear under the robust ABC-R procedure, which again signals that the classical LMM posterior is driven by few extreme observations (see Sect. 2).

### 6 Final remarks

Former approaches on the use of robust unbiased estimating functions in Bayesian inference have been focused mainly on the quasi- and the empirical likelihoods as replacement of the full likelihood in Bayes’ formula (see, e.g., Greco et al., 2008). However, quasi-likelihood is available only for scalar parameters of interest while empirical likelihood can be unstable for small sample sizes.
In this respect, the present work offers a contribution by providing an alternative approach for building posterior distributions from robust estimating functions using simulated data from the central model and comparing them with the one actually observed, in an ABC fashion. Such a comparison is done through the chosen robust estimating function and this permits us to condition upon a robust subset of the data.

The method shares some limitations of the ABC setting. Firstly, a tolerance threshold has to be chosen. We fix this tolerance in order to give a pre-specified but small acceptance ratio, as frequently done in the ABC literature. Secondly, Bayesian hypothesis testing via BFs or posterior model probabilities is generally infeasible since the ABC model selection algorithm (Robert et al., 2011) requires ad hoc summary statistics (Marin et al., 2014), which are generally very difficult to find. The full Bayesian significance testing setting (Pereira et al., 2008) is another practical and theoretically sound approach to Bayesian testing. Furthermore, it is easy to implement since for testing scalar parameters it requires only marginal posterior densities, which can be readily built by, say, kernel density estimation.

Motivated by the GRP94 dataset, we considered only two-component nested LMM, but more complex models can be fitted since the estimating equations (10a)-(10b) are very general (see Richardson, 1997). For instance, it is possible to deal with models with many random effects or even with robustness with respect to the design matrix. Moreover, the proposed ABC-R method works with any unbiased estimating equations, hence $S$- and $MM$-estimating functions can be readily used.

**Acknowledgement**

This work was partially supported by University of Padova (Progetti di Ricerca di Ateneo 2015, CPDA153257) and by PRIN 2015 (grant 2015EASZFS_003).

**References**

Adimari, G. and Ventura, L. (2002a) Quasi-likelihood from M-estimators: A numerical comparison with empirical likelihood. *Statistical Methods and Applications*, **11**, 175–185.
— (2002b) Quasi-profile log likelihoods for unbiased estimating functions. *Annals of the Institute of Statistical Mathematics*, 54, 235–244.

Agostinelli, C. and Greco, L. (2013) A weighted strategy to handle likelihood uncertainty in Bayesian inference. 28, 319–339.

Andrade, J. A. A. and O’Hagan, A. (2006) Bayesian robustness modeling using regularly varying distributions. 1, 169–188.

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998) Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85, 549–559.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67, 1–48.

Beaumont, M. A., Zhang, W. and Balding, D. J. (2002) Approximate Bayesian computation in population genetics. *Genetics*, 162, 2025–2035.

Berger, J. O. (1994) An overview of robust bayesian analysis. *Test*, 3, 5–124.

Blum, M. G. B. (2010) Approximate Bayesian computation: a nonparametric perspective. 105, 1178–1187.

Boos, D. D. and Monahan, J. F. (1986) Bootstrap methods using prior information. *Biometrika*, 73, 77–83.

Cabras, S., Castellanos Nueda, M. E. and Ruli, E. (2015) Approximate Bayesian computation by modelling summary statistics in a quasi-likelihood framework. *Bayesian Analysis*, 10, 411–439.

Cattelan, M. and Sartori, N. (2015) Empirical and simulated adjustments of composite likelihood ratio statistics. *Journal of Statistical Computation and Simulation*, 86, 1056–1067.

Copt, S. and Heritier, S. (2006) Robust MM-estimation and inference in mixed linear models. *Econometrics Department Research Report: 2006.01, University of Geneva*, Geneva, Switzerland. [http://www.unige.ch/ses/metri/cahiers/2006_01.pdf](http://www.unige.ch/ses/metri/cahiers/2006_01.pdf).
Copt, S. and Victoria-Feser, M.-P. (2006) High-breakdown inference for mixed linear models. 101, 292–300.

Davison, A. C., Hinkley, D. V. and Worton, B. J. (1992) Bootstrap likelihoods. *Biometrika*, 79, 113–130.

Dawid, A. P. and Musio, M. (2014) Theory and applications of proper scoring rules. *Metron*, 72, 169–183.

Dawid, A. P., Musio, M. and Ventura, L. (2016) Minimum scoring rule inference. 43, 123–138.

Doksum, K. A. and Lo, A. Y. (1990) Consistent and robust bayes procedures for location based on partial information. 18, 443–453.

Drovandi, C. C., Pettitt, A. N. and Faddy, M. J. (2015) Bayesian indirect inference using a parametric auxiliary model. 30, 72–95.

Farcomeni, A. and Greco, L. (2015) *Robust methods for data reduction*. CRC press.

Farcomeni, A. and Ventura, L. (2012) An overview of robust methods in medical research. *Statistical Methods in Medical Research*, 21, 111–133.

Fearnhead, P. and Prangle, D. (2012) Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation (with Discussion). 74, 419–474.

Gelman, A. (2006) Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian analysis*, 1, 515–534.

Gleim, A. and Pigrosch, C. (2013) Approximate Bayesian computation with inderect summary statistics. *Technical report*, University of Bonn, Germany.

Greco, L., Racugno, W. and Ventura, L. (2008) Robust likelihood functions in Bayesian inference. 138, 1258–1270.

Grünwald, P. and van Ommen, T. (2014) Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *arXiv preprint arXiv:1412.3730*. 
Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986) Robust Statistics. The approach based on influence function. Chichester, UK: John Wiley & Sons.

Heritier, S., Cantoni, E., Copt, S. and Victoria-Feser, M. P. (2011) Robust Methods in Biostatistics. Chichester, UK: John Wiley & Sons.

Huber, P. J. (1981) Robust Statistics. New York: John Wiley & Sons.

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

Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989) Robust statistical modeling using the t distribution. Journal of the American Statistical Association, 84, 881–896.

Lavine, M. (1995) On an approximate likelihood for quantiles. Biometrika, 82, 220–222.

Lazar, N. A. (2003) Bayesian empirical likelihood. Biometrika, 90, 319–326.

Lewis, J. R., MacEachern, S. N. and Lee, Y. (2014) Bayesian restricted likelihood. Technical report No. 878, The Ohio State University, USA.

Marin, J.-M., Pillai, N. S., Robert, C. P. and Rousseau, J. (2014) Relevant statistics for Bayesian model choice. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76, 833–859.

Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003) Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences, 100, 15324–15328.

Markatou, M., Basu, A. and Lindsay, B. G. (1998) Weighted likelihood equations with bootstrap root search. Journal of the American Statistical Association, 93, 740–750.

McCullagh, P. (1991) Quasi-likelihood and estimating functions. In Statistical Theory and Modelling (eds. D. Hinkley, N. Reid and E. Snell), 265–286. Chapman and Hall: London.

Miller, J. W. and Dunson, D. B. (2015) Robust Bayesian inference via coarsening. arXiv preprint arXiv:1506.06101.
Monti, A. C. and Ronchetti, E. (1993) On the relationship between empirical likelihood and empirical saddlepoint approximation for multivariate M-estimators. *Biometrika*, 80, 329–338.

Owen, A. B. (2001) *Empirical Likelihood*. Boca Ranton, Florida: Chapman and Hall/CRC.

Pace, L. and Salvan, A. (1997) *Principles of Statistical Inference*. Singapore: World Scientific.

Patterson, H. D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, 545–554.

Pereira, C. A. d. B. and Stern, J. M. (1999) Evidence and credibility: a full Bayesian test of precise hypothesis. *Entropy*, 1, 14–115.

Pereira, C. A. d. B., Stern, J. M. and Wechsler, S. (2008) Can a significance test be genuinely Bayesian? *3*, 79–100.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. (1999) Population growth of human y chromosomes: a study of y chromosome microsatellites. *16*, 1791–1798.

Richardson, A. M. (1997) Bounded influence estimation in the mixed linear model. *92*, 154–161.

Richardson, A. M. and Welsh, A. H. (1995) Robust restricted maximum likelihood in mixed linear models. *Biometrics*, 51, 1429–1439.

Ríos Insua, D. and Ruggeri, F. (2000) *Robust Bayesian Analysis*. Lecture Notes in Statistics 152. New York: Springer.

Robert, C. P., Cornuet, J.-M., Marin, J.-M. and Pillai, N. S. (2011) Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, 108, 15112–15117.

Ronchetti, E. M. and Welsh, A. H. (1994) Empirical saddlepoint approximations for multivariate M-estimators. *56*, 313–326.
Rousseeuw, P. J. and Leroy, A. M. (2003) *Robust Regression and Outlier Detection*. Hoboken, New Jersey: John Wiley & Sons.

Rubin, D. B. (1981) The Bayesian bootstrap. *The Annals of Statistics*, 9, 130–134.

Ruggeri, F., Ríos Insua, D. and Berger, J. O. (2005) Robust Bayesian analysis. In *Bayesian Thinking: Modeling and Computation. Handbook of Statistics*, vol. 25, 623–663. Amsterdam: Elsevier.

Ruli, E., Sartori, N. and Ventura, L. (2016) Approximate Bayesian computation with composite score functions. *26*, 679–692.

Soubeyrand, S., Carpentier, F., Guiton, F. and Klein, E. K. (2013) Approximate Bayesian computation with functional statistics. *12*, 17–37.

Soubeyrand, S. and Haon-Lasportes, E. (2015) Weak convergence of posteriors conditional on maximum pseudo-likelihood estimates and implications in abc. *107*, 84–92.

Tavaré, S., Balding, D. J., Griffiths, R. C. and Donnelly, P. (1997) Inferring coalescence times from dna sequence data. *Genetics*, 145, 505–518.

Tramentozzi, E., Ruli, E., Angriman, I., Angriman, I., Bardini, R., Bardini, R., Campora, M., Campora, M., Guzzardo, V., Guzzardo, V., Zamarchi, R., Zamarchi, R., Rossi, E., Rossi, E., Rugge, M., Rugge, M., Finotti, P. and Finotti, P. (2016) Grp94 in complexes with IgG is a soluble diagnostic marker of gastrointestinal tumors and displays immune-stimulating activity on peripheral blood immune cells. *Oncotarget*, 7, 72923–72940.

Tsou, T.-S. and Royall, R. M. (1995) Robust likelihoods. *Journal of the American statistical Association*, 90, 316–320.

Ventura, L., Cabras, S. and Racugno, W. (2010) Default prior distributions from quasi- and quasi-profile likelihoods. *43*, 2937–2942.

Ventura, L. and Racugno, W. (2016) Pseudo-likelihoods for Bayesian inference. In *Selected Papers of the Scientific Meetings of the Italian Statistical Society*, Lecture Notes–Monograph Series, vol. 29, 205–222. Springer-Verlag Italia.
Watson, J. and Holmes, C. (2016) Approximate models and robust decisions. *Statistical Science, 31*, 465–489.

Welsh, A. H. and Richardson, A. M. (1997) Approaches to the robust estimation of mixed models. In *Handbook of Statistics* (eds. G. S. Maddala and C. R. Rao), vol. 15, 343–384. Amsterdam: Elsevier Science.