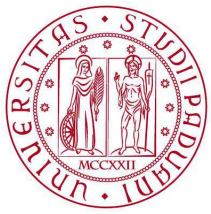


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXVI

Recent Advances in Approximate Bayesian Computation Methods

Direttore della Scuola: Ch.ma Prof.ssa Monica Chiogna

Supervisore: Ch.ma Prof.ssa Laura Ventura

Co-supervisore: Ch.mo Prof. Nicola Sartori

31 Gennaio 2014

Dottorando: Erlis Ruli

To Reiz

Acknowledgements

I cannot express my gratitude enough to Laura & Walter for their constant support and encouragement. Thanks to their sincere advices, not only statistical, I was able to start and successfully finish this PhD adventure.

The completion of this thesis would not have been possible without Nicola's guidance, from whom I had also the opportunity to learn how to perform research. I am indebted to Stefano and Maria Eugenia, for inviting me at the Universidad Carlos III de Madrid, where I had the opportunity to exploit new interesting topics and make noticeable progress on my PhD thesis. Special thanks also go to Mike and his colleagues, for the pleasant time we had talking, and having cerveza with tapas.

I am grateful to Christian P. Robert and Nicolas Chopin for the excellent learning opportunities I had during my visiting at CREST. Many of the ideas developed in the thesis came to my mind during my time spent at CREST, thanks to the feedback provided by Christian and Nicolas. Thanks also to Marco and Sofia for their warm hospitality.

I wish to thank Anthony C. Davison and Ioannis Kosmidis for their useful comments and questions that improved the thesis.

I am also grateful to the Department of Statistical Sciences for giving me this opportunity, and to the administrative and technical staff for their support over the past three years.

Thanks to Akram, Darda, Ivan, Lorenzo, Luca, Roberta and Shireen for the good times we had together and the opportunity to exchange ideas.

Finally, thanks to Maria, my objective prior, for the beautiful baby bump, for her patience and her unconditional support during all these years.

Padova, Erlis Ruli

22 January 2014

Abstract

The Bayesian approach to statistical inference in fundamentally probabilistic. Exploiting the internal consistency of the probability framework, the posterior distribution extracts the relevant information in the data, and provides a complete and coherent summary of post data uncertainty. However, summarising the posterior distribution often requires the calculation of awkward multidimensional integrals. A further complication with the Bayesian approach arises when the likelihood functions is unavailable. In this respect, promising advances have been made by theory of Approximate Bayesian Computations (ABC).

This thesis focuses on computational methods for the approximation of posterior distributions, and it discusses six original contributions. The first contribution concerns the approximation of marginal posterior distributions for scalar parameters. By combining higher-order tail area approximations with the inverse transform sampling, we define the HOTA algorithm which draws independent random samples from the approximate marginal posterior. The second discusses the HOTA algorithm with pseudo-posterior distributions, *e.g.*, posterior distributions obtained by the combination of a pseudo-likelihood with a prior within Bayes' rule. The third contribution extends the use of tail-area approximations to context with multidimensional parameters, and proposes a method which gives approximate Bayesian credible regions with good sampling coverage properties. The forth presents an improved Laplace approximation which can be used for computing marginal likelihoods. The fifth contribution discusses a model-based procedure for choosing good summary statistics for ABC, by using composite score functions. Lastly, the sixth contribution discusses the choice of a default proposal distribution for ABC that is based on the notion of quasi-likelihood.

Sommario

L'approccio bayesiano all'inferenza statistica è fondamentalmente probabilistico. Attraverso il calcolo delle probabilità, la distribuzione a posteriori estrae l'informazione rilevante offerta dai dati e produce una descrizione completa e coerente dell'incertezza condizionatamente ai dati osservati. Tuttavia, la descrizione della distribuzione a posteriori spesso richiede il calcolo di integrali multivariati e complicati. Un'ulteriore difficoltà dell'approccio bayesiano è legata alla funzione di verosimiglianza e nasce quando quest'ultima è matematicamento o computazionalmente intrattabile. In questa direzione, notevoli sviluppi sono stati ottenuti attraverso la cosiddetta teoria di computazioni bayesiane approssimate (ABC).

Questa tesi si focalizza su metodi computazionali per l'approssimazione della distribuzione a posteriori e propone sei contributi originali. Il primo contributo concerne l'approssimazione della distribuzione a posteriori marginale per un parametro scalare. Combinando l'approssimazione di ordine superiore per aree nelle code con il metodo della simulazione per inversione, si ottiene l'algorimo denominato HOTA, il quale può essere usato per simulare in modo indipendente da un'approssimazione della distribuzione a posteriori. Il secondo contributo si propone di estendere l'uso dell'algoritmo HOTA in contesti di distribuzioni pseudo-a posteriori, ovvero distribuzioni a posteriori ottenute attraverso la combinazione di una opportuna pseudo-verosimiglianza con una distribuzione a priori, tramite il teorema di Bayes. Il terzo contributo estende l'uso dell'approssimazione per aree nelle code in contesti con parametri multidimensionali e propone un metodo per ottenere delle regioni di credibilità con buone proprietà di copertura frequentista. Il quarto contributo presenta un'approssimazione di Laplace di terzo ordine per il calcolo della verosimiglianza marginale. Il quinto contributo si focalizza sulla scelta delle statistiche descrittive per i metodi ABC e propone un'approccio basato sulla funzione composita punteggio, per la scelta di tali statistiche. Infine, l'ultimo contributo si focalizza sulla scelta automatica di una distribuzione di proposta per algoritmi ABC, dove la procedura di derivazione di tale distribuzione è basata sulla nozione della quasi-verosimiglianza.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

The Bayesian approach to statistical inference is fundamentally probabilistic. The relationships between all the unknowns and the data are described by the likelihood function times the prior distribution on the unknowns. Straightforward application of Bayes' theorem provides the conditional probability distribution of the unknowns given the data, *i.e.* the posterior distribution. Beyond the specification of the likelihood and the prior, the Bayesian approach is automatic. Exploiting the internal consistency of the probability framework, the posterior distribution extracts the relevant information in the data and provides a complete and coherent summary of post data uncertainty (Bernardo & Smith, 1994). Using the posterior to solve specific inference and decision problems is then straightforward, at least in principle.

In practice, the Bayesian approach typically faces two major challenges: the specification of the prior and the calculation of the posterior distribution. Sometimes prior elicitation can be performed by adopting a default approach, such as that outlined by Jeffreys (see Kass & Wasserman, 1996). Moreover, when a probability matching between Bayesian and frequentist procedures is of interest, matching priors can be considered (see, *e.g.*, Ventura *et al.*, 2009, and references therein) or when historical data are available they can be used for constructing prior distributions in a subjective way. Posterior computations can be performed analytically only in few instances, *e.g.* models with conjugate priors, hence in general approximation methods are needed. There exist many useful posterior computation methods, which range from asymptotic methods based on the Laplace approximation (Tierney & Kadane, 1986) to Monte Carlo methods (see, *e.g.*, Chen *et al.*, 2000). Asymptotic methods, are typically computationally faster, easier to implement, and do not require tuning. Monte Carlo methods are typically

more involved, and require more attention from the practitioner, especially Monte Carlo methods based on Markov chains.

A further complication arises as far as the computation of the posterior normalizing constant is concerned. The complication arises because, most of the posterior simulation methods typically sidestep the normalizing constant, hence its computation needs further efforts (see, *e.g.*, Chen *et al.*, 2000, Ch. 5). The normalizing constants are used to compute posterior model probabilities, which are useful for Bayesian model selection. Selection strategies based on posterior model probabilities can be motivated via a decision theoretic framework where the goal is to maximize the expected utility (Bernardo & Smith, 1994, Ch. 6; Robert, 2007, Ch. 7).

Another major challenge to the Bayesian approach (as well as to all likelihood-based inferential settings) arises when the likelihood function is analytically or computationally intractable. This is often the case in complex models, *e.g.* models with complicated dependence structures, with many latent variables or semi-parametric models. At present, there are mainly two ways to deal with complex models. The first one is to use suitable approximate likelihoods with similar properties to the full likelihood, called pseudo-likelihoods (see, *e.g.*, Pace & Salvan, 1997, Ch. 4; Varin *et al.*, 2011), as a surrogate of the full likelihood in the Bayes' rule. Although this approach cannot always be considered orthodox in a Bayesian setting, the use of approximate likelihoods is nowadays widely shared, and several papers are devoted to Bayesian interpretation and applications of some well known pseudo-likelihoods, such as the composite likelihood and the partial likelihood. For the use of composite likelihoods in a Bayesian approach see Smith & Stephenson (2009), Pauli *et al.* (2011) and Ribatet *et al.* (2012). The second approach is to resort to techniques known as likelihood-free or Approximate Bayesian Computations (ABC) methods (see Marin *et al.*, 2012, for a review).

1.2 Main contributions of the thesis

This thesis focuses on computational methods for approximating posterior distributions and related quantities, and it proposes six original contributions.

1. For models in which the likelihood function is available and regular, and using higher-order asymptotics, we develop the HOTA algorithm which gives independent samples from the approximate marginal posterior distribution for scalar parameters. It is a combination of the higher-order tail area approximation (see, *e.g.*, Reid, 2003) with the inverse transform sampling. An advantage of HOTA with

respect to other methods, such as MCMC, is that it is typically very fast and does not require convergence checks. An `R` package is currently under development, which implements the HOTA algorithm. The method is discussed in Section 3.1.

2. In the context of models where the likelihood function is unavailable and is replaced by a suitable pseudo-likelihood, we derive a higher-order tail area approximation for pseudo-posterior distributions, *e.g.* posterior distributions obtained from the combination of pseudo-likelihoods and priors in the Bayes' rule. We study the accuracy of this formula by means of practical examples and discuss its use via the HOTA algorithm. This contribution is illustrated in Section 3.2.

3. In Section 3.3, using higher-order asymptotics for a multidimensional parameter of interest, we derive credible sets which can be interpreted as an extension of the Bayesian equi-tailed credible intervals to a multivariate setting. We show, by means of practical examples, that the approximate Bayesian credible sets have accurate posterior probability contents and good sampling properties.

4. In Section 3.4, we discuss how the Laplace approximation for marginal posterior distributions (Tierney & Kadane, 1986) can be combined with the marginal likelihood approach of Chib (1995) in order to obtain posterior normalizing constants which are accurate to $O(n^{-3/2})$, where n is the sample size. The proposed method is more accurate than the usual Laplace approximation for posterior normalizing constants and requires only numerical integration.

5. Our first contribution to ABC theory concerns the choice of the summary statistics which is still an open problem. In particular, we propose to use the composite score function (see, *e.g.*, Varin *et al.*, 2011) evaluated at a fixed parameter value, as a summary of the data. This gives rise to a new algorithm, called the ABC-cs algorithm. The advantage of this method is that it automatically defines a statistic which incorporates useful characteristics of the complex model. The method is illustrated with several examples in Section 4.1.

6. Lastly, in Section 4.2, we use the theory of quasi-likelihoods (see, *e.g.*, Pace & Salvan, 1997, Ch. 4) to construct a default proposal distribution for MCMC or importance sampling algorithms within ABC. Given an estimated binding function, *e.g.* a monotone regression between the summary statistics and the parameter, along with the related Jacobian, our proposal produces candidate values in the space of the summary statistics which are then transformed, via the binding function, in terms of parameter values. The method is illustrated by an example.

Chapter 2

Bayesian Approximation Methods

2.1 Preamble

The approximation or exploration of the posterior distribution is one of the fundamental difficulties with the Bayesian approach. In contexts where the likelihood is available analytically or numerically and a prior distribution is given, we must deal with the problem of summarizing the posterior density, which in general consists on computing awkward multidimensional integrals. A further difficulty with both the Bayesian and the frequentist approach arises in complex models, where the likelihood function may be difficult or even impossible to evaluate.

This chapter gives a short overview of some of the most popular Bayesian approximation methods, useful to summarize posterior distributions and to compute posterior normalizing constants. The chapter is structured as follows. Section 2.2 states the problem of posterior computation, along with notation used throughout the thesis. Section 2.3 describes asymptotic and higher-order approximation methods. Section 2.4 describes Monte Carlo methods based on Markov chains and importance sampling techniques. Section 2.5 describes pseudo-likelihoods and likelihood-free techniques useful to deal with complex models.

2.2 Context

Consider a parametric statistical model with probability density function $p(y;\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^d$, and let $L(\theta) = L(\theta; y) = \prod_{i=1}^{n} p(y_i;\theta)$ be the likelihood function of θ based on a random sample $y = (y_1, \ldots, y_n)^T$ of size n.

Bayesian inference on θ is carried out through the posterior density

$$\pi(\theta|y) \;=\; \frac{L(\theta)\pi(\theta)}{p(y)}\,, \tag{2.1}$$

where $\pi(\theta)$ is a prior distribution for θ and

$$p(y) \;=\; \int_{\Theta} L(\theta)\pi(\theta)\,d\theta \tag{2.2}$$

is the normalizing constant, often called the marginal likelihood or model evidence, *i.e.* the marginal density of the data under the assumed model and prior.

From the posterior density (2.1), and using standard probability marginalization techniques we can in principle straightforwardly calculate any particular univariate, bivariate, etc., summary in the form of densities, contours or moments, as required. However, the calculation of the joint posterior density of θ, and the required marginalization and moment summaries, rests on the ability to compute high-dimensional integrals, which in practice may be high-dimensional and tedious.

For instance, the posterior expectation of a given function $g(\theta)$, defined as

$$\begin{aligned} E_{\pi}(g) &= \int_{\Theta} g(\theta)\pi(\theta|y)\,d\theta \\ &= \frac{\int_{\Theta} g(\theta)L(\theta)\pi(\theta)\,d\theta}{\int_{\Theta} L(\theta)\pi(\theta)\,d\theta}, \end{aligned} \tag{2.3}$$

involves a ratio of d-dimensional integrals. In the particular case of $g(\theta) = p(z;\theta)$, with z being a future observation, the expectation is known as posterior predictive density, *e.g.* the posterior density of z given the observed data y.

Another practical scenario is when $\theta = (\psi, \lambda)$, where ψ is a p-dimensional parameter of interest and λ the nuisance parameter of dimension $d - p$. Then, Bayesian inference on ψ can be based on the marginal posterior density

$$\begin{aligned} \pi(\psi|y) &= \int_{\Lambda} \pi(\psi, \lambda|y)\,d\lambda \\ &= \frac{\int_{\Lambda} L(\psi, \lambda)\pi(\psi, \lambda)\,d\lambda}{\int_{\Theta} L(\theta)\pi(\theta)\,d\theta}, \end{aligned} \tag{2.4}$$

which is again a ratio of multivariate integrals.

A further complication arises in the computation of the posterior normalizing constant (2.2) because most of the computational methods used to simulate from (2.1) typically sidestep the normalizing constant. Hence, its computation needs further efforts (see, *e.g.*, Chen *et al.*, 2000, Ch. 5).

The normalizing constant is typically used for Bayesian model selection (see, *e.g.*, Robert, 2007, Ch. 7). In particular, suppose it is of interest to select the "best" model from a set of candidate models $(M_1, \ldots, M_K)$. For each model M_k, let $p_k(y|\theta_k)$ be the associated probability density with $\theta_k \in \Theta_k \subseteq \mathbb{R}^{d_k}$, and denote by $L(\theta_k; M_k) = L(\theta_k; M_k, y)$ and $\pi(\theta_k|M_k)$ the associated likelihood and prior, respectively $(k = 1, \ldots, K)$. The posterior distribution of θ_k for model M_k is

$$\pi(\theta_k|M_k, y) = \frac{L(\theta_k; M_k)\pi(\theta_k|M_k)}{p(y|M_k)}, \tag{2.5}$$

where $p(y|M_k) = \int_{\Theta_k} L(\theta_k; M_k)\pi(\theta_k|M_k)\, d\theta_k$ is the normalizing constant of $\pi(\theta_k|y, M_k)$ $(k = 1, \ldots, K)$. Statements about posterior model probabilities can based on

$$p(M_k|y) = \pi(M_k)p(y|M_k) / \sum_{k=1}^{K} \pi(M_k)p(y|M_k),$$

where $\pi(M_k)$ is the prior for model M_k $(k = 1, \ldots, K)$. The posterior model probability distribution $\{p(M_1|y), \ldots, p(M_K|y)\}$ is a fundamental object of interest in model selection. Insofar as the priors $\pi(\theta_k|M_k)$ and $\pi(M_k)$ provide an initial representation of model uncertainty, the posterior model probability summarizes all the relevant information in the data y and provides a complete post-data representation of model uncertainty (Chipman *et al.*, 2001). By treating $\pi(M_k|y)$ as a measure of the "truth" of model M_k, a natural and simple strategy for model selection is to choose the most probable M_k, the one for which $\pi(M_k|y)$ is largest. Alternatively, one might prefer to report a set of high posterior models along with their probabilities to convey the model uncertainty. Selection strategies based on the posterior model probability can be motivated via a decision theoretic framework, where the goal is to maximize the expected utility (see, Robert, 2007, Ch. 7, and Bernardo & Smith, 1994).

Posterior normalizing constants can be used also to compare models via Bayes factors (BFs), given by

$$B_{ij} = \frac{p(y|M_i)}{p(y|M_j)}, \quad i \neq j = 1, \ldots, K, \tag{2.6}$$

which give the relative evidence that is provided by the data in favour of model M_i compared with model M_j (see, *e.g.*, Kass & Raftery, 1995). Notice that for the purposes of model choice, priors $\pi(\theta_k|M_k)$ must necessarily be proper, since improper priors give BFs which are not well determined (Kass & Raftery, 1995). This is because, for instance, the marginal likelihood of a model with prior $\pi(\theta) \propto c$, with $c > 1$, is c times higher than the marginal likelihood of the same model based on the same data but with prior $\pi(\theta) \propto 1$. See also Lahiri (2001) for recent developments on the use of BFs for model selection, and Lavine & Schervish (1999) for a critical view on BFs.

In few special cases, the integrals involved in (2.2), (2.3) and (2.4) can be computed exactly. Otherwise approximation methods must be used.

2.3 Asymptotic approximations

Posterior computations based on asymptotic arguments are perhaps the oldest computational methods (see, *e.g.* Lindley, 1965). The overall idea of these methods is that, for a large sample size, the likelihood will be roughly normal and dominated by a single mode. Hence, various approximation methods can be constructed by considering suitable Taylor expansions and integrations over normal-type functions. This is the idea of Laplace's method of integration (see, *e.g.*, Tierney & Kadane, 1986). An essential requirement of this type of approximation is that the log-posterior or the log-likelihood of the model must be a smooth function of θ, with a unique mode; see Kass *et al.* (1990) for a detailed exposition of the regularity conditions.

Asymptotic methods give analytical approximations, *e.g.* manageable and fixed functions of the parameters and the data, and this is both an advantage and a disadvantage. The advantage is that the analytical expressions obtained from these asymptotic methods are easy to handle and to program. Typically, asymptotic methods require only maximization and differentiation routines. On the other hand, the accuracy of the approximations depends on the "degree" of the normality of the likelihood or of the posterior, which in turn is essentially governed by the fixed sample size. Moreover the approximation error of these asymptotic methods is guaranteed to be zero as $n \to \infty$, but finite sample error bounds are generally unavailable.

However, the accuracy of the asymptotic methods may be sufficiently high for many practical applications, especially in cases when stochastic or Monte Carlo approximations converge too slowly to be useful (see, *e.g.*, Rue *et al.*, 2009). Moreover, asymptotic methods can be used in conjunction with Monte Carlo methods (see, *e.g.*, Guihenneuc-Jouyaux & Rousseau, 2005; Kharroubi & Sweeting, 2010).

2.3.1 Normal approximations

The normal approximation is perhaps the simplest approach among all posterior computation methods (see, *e.g.*, O'Hagan & Forster, 2004, Ch. 8). Let $H(\theta) =$

$\pi(\theta)\exp\{\ell(\theta)\}$ be the posterior kernel, where $\ell(\theta) = \log L(\theta)$ is the log-likelihood function. The normal approximation can be derived as follows. Consider the Taylor expansion of $h(\theta) = \log H(\theta)$ about the posterior mode $\tilde{\theta} = \arg\max_{\theta\in\Theta} H(\theta)$, *i.e.*

$$h(\theta) = h(\tilde{\theta}) + \frac{1}{2}(\theta - \tilde{\theta})^T V(\tilde{\theta})(\theta - \tilde{\theta}) + R,$$

where $V(\tilde{\theta}) = -\partial^2 h(\theta)/\partial\theta\partial\theta^T|_{\theta=\tilde{\theta}}$ is the posterior information matrix, *e.g.* the Hessian matrix of the negative log-posterior evaluated at $\tilde{\theta}$, and R is the remainder of order $O(n^{-1/2})$. After substituting the expanded $h(\theta)$ in (2.1), and integrating the expansion in the denominator, we get that the posterior distribution is approximately normal, *i.e.* $\theta|y \overset{a}{\sim} N(\tilde{\theta}, \Sigma(\tilde{\theta}))$, centred at the posterior mode with variance-covariance matrix $\Sigma(\tilde{\theta}) = V(\tilde{\theta})^{-1}$, where the symbol "$\overset{a}{\sim}$" means asymptotically distributed for $n \to \infty$ and the error is of order $O(n^{-1/2})$ (see, *e.g.* , Lindley, 1965, and Walker, 1969, for a precise statement of this result).

The normalizing constant of the normal posterior is readily found to be

$$\begin{aligned} p(y) &= (2\pi)^{d/2} H(\tilde{\theta})|V(\tilde{\theta})|^{-1/2}\{1 + O(n^{-1})\} \\ &= p_L(y)\{1 + O(n^{-1})\} \end{aligned} \tag{2.7}$$

and any posterior summary is immediately obtained by using known results about the multivariate normal distribution. For instance, the posterior mean is $\tilde{\theta}$, whereas the marginal posterior of θ_i is $N(\tilde{\theta}_i, \Sigma_{ii}(\tilde{\theta}))$, where $\Sigma_{ii}(\tilde{\theta})$ denotes the (i,i) block of the matrix $\Sigma(\tilde{\theta})$, $i = 1, \ldots, d$.

Another asymptotically equivalent version to the normal approximation can be computed by taking a Taylor expansion of the log-likelihood function around the maximum likelihood estimate (MLE) $\hat{\theta}$. However, although the theoretical approximation error is still of order $O(n^{-1/2})$ (see, *e.g.* Reid, 1996), the approximation could be less accurate, especially in small samples (O'Hagan & Forster, 2004, p. 214) or when the MLE and the posterior mode are very different.

2.3.2 Higher-order approximations

The normal approximation is sufficiently accurate if the posterior distribution is approximately quadratic. Therefore, this approximation applied to asymmetric or skewed posteriors can be severely inaccurate. A first way to improve the normal approximation is to include higher-order derivatives of $h(\theta)$ in the Taylor expansions (see Lindley, 1961, 1980). However, this route to higher-order asymptotics for posterior approximations can

be impracticable, because log-likelihood or log-posterior derivatives of order higher than the second can be cumbersome to compute in practice, even numerically.

The evaluation of $\int_\Theta \exp\{h(\theta)\}\, d\theta$ using Taylor expansions about $\tilde{\theta}$ and integrating term by term is an application of Laplace's method of integration; see, *e.g.* DeBruijn (1961, Ch. 4), Barndorff-Nielsen & Cox (1989, Ch. 6), and Tierney & Kadane (1986).

A higher-order approximation to the posterior distribution, which requires only first and second order log-posterior derivatives, can obtained by applying the Laplace approximation to the denominator of (2.1), but not expanding the numerator, namely

$$\pi(\theta|y) \;=\; (2\pi)^{-d/2}|V(\tilde{\theta})|^{1/2}\{h(\theta) - h(\tilde{\theta})\}\{1 + O(n^{-1})\}. \tag{2.8}$$

Another asymptotically equivalent version of (2.8) can be obtained by expanding the log-likelihood $\ell(\theta)$ around the MLE $\hat{\theta}$ and leaving the prior unchanged, which gives (see, *e.g.*, Reid, 1996, 2003)

$$\pi(\theta|y) \;=\; (2\pi)^{-d/2}|j(\hat{\theta})|^{1/2}\exp\{\ell(\theta) - \ell(\hat{\theta})\}\frac{\pi(\theta)}{\pi(\hat{\theta})}\{1 + O(n^{-1})\}, \tag{2.9}$$

where $j(\hat{\theta}) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^T|_{\theta=\hat{\theta}}$ is the observed Fisher information matrix.

When $\theta = (\psi, \lambda)$, with ψ the parameter of interest, then the Laplace approximation to the marginal posterior (2.4) can be obtained as follows. Let $\tilde{\theta} = (\tilde{\psi}, \tilde{\lambda})$ and let $\tilde{\theta}_\psi = (\psi, \tilde{\lambda}_\psi)$, where $\tilde{\lambda}_\psi$ is the posterior mode with ψ fixed, *i.e.* $\tilde{\lambda}_\psi = \arg\max_{\lambda\in\Lambda} h(\psi, \lambda)$. Expand both the numerator of (2.4) about $\tilde{\lambda}_\psi$ and the denominator about $\tilde{\theta}$, up to the quadratic term. Then, after integrating term by term both the expansions in the numerator and the denominator, we obtain (Tierney & Kadane, 1986)

$$\pi(\psi|y) \;=\; (2\pi)^{-p/2}\exp\{h(\tilde{\theta}_\psi) - h(\tilde{\theta})\}\left\{\frac{|V(\tilde{\theta})|}{|V_{\lambda\lambda}(\tilde{\theta}_\psi)|}\right\}^{1/2}\{1 + O(n^{-3/2})\}, \tag{2.10}$$

where $\tilde{V}_{\lambda\lambda}(\tilde{\theta}_\psi) = V_{\lambda\lambda}(\theta)|_{\theta=\tilde{\theta}_\psi}$. This approximation, after numerical renormalization, tends to be very accurate in practice (Tierney & Kadane, 1986).

The MLE-based version of (2.10) is given Reid (2003) and is

$$\pi(\psi|y) \;=\; \frac{|j_p(\hat{\psi})|^{1/2}}{(2\pi)^{p/2}}\,\exp\{\ell_p(\psi) - \ell_p(\hat{\psi})\}\left\{\frac{|j_{\lambda\lambda}(\hat{\theta})|}{|j_{\lambda\lambda}(\hat{\theta}_\psi)|}\right\}^{1/2}\frac{\pi(\hat{\theta}_\psi)}{\pi(\hat{\theta})}\{1 + O(n^{-3/2})\}, \tag{2.11}$$

where $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$, $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$, with $\hat{\lambda}_\psi$ the constrained MLE of λ for fixed ψ, $\ell_p(\psi) = \ell(\psi, \hat{\lambda}_\psi)$ is the profile log-likelihood, and $j_p(\psi) = -\partial^2\ell_p(\psi)/\partial\psi^T$ the profile information.

Posterior moments for scalar components can be obtained by integrating the renormalized versions of (2.10) or (2.11), whereas an expression for posterior expectations of the complete parameter vector can be found in Tierney & Kadane (1986).

The Laplace approximation (2.10) is also the basis for the derivation of accurate posterior tail areas for scalar parameters. In particular, let ψ be a scalar parameter of interest. Then it can be shown that (see Davison, 2003, Ch. 11)

$$\begin{aligned}
\mathrm{P}(\psi < \psi_0|y) &= \int_{-\infty}^{\psi_0} \pi(\psi|y)\, d\psi \\
&= \int_{-\infty}^{\psi_0} (2\pi)^{-1/2} \exp\{h(\tilde\theta_\psi) - h(\tilde\theta)\} \left\{ \frac{|V(\tilde\theta)|}{|V_{\lambda\lambda}(\tilde\theta_\psi)|} \right\}^{1/2} \{1 + O(n^{-3/2})\}\, d\psi \\
&= \Phi\left\{ r_p^B(\psi_0) + \frac{1}{r_p^B(\psi_0)} \log \frac{q_B(\psi_0)}{r_p^B(\psi_0)} \right\} \{1 + O(n^{-3/2})\} \\
&= \Phi\{r_B^\star(\psi_0)\}\{1 + O(n^{-3/2})\}, \qquad (2.12)
\end{aligned}$$

where $r_p^B(\psi) = \mathrm{sign}(\psi - \tilde\psi)\{2[h(\tilde\theta) - h(\tilde\theta_\psi)]\}^{1/2}$, $r_B^\star(\psi) = r_p^B(\psi) + r_p^B(\psi)^{-1} \log\{q_B(\psi)/r_p^B(\psi)\}$ and

$$q_B(\psi) = -h_\psi(\tilde\theta_\psi)\{|V_{\lambda\lambda}(\tilde\theta_\psi)|/|V(\tilde\theta)|\}^{1/2},$$

with $h_\psi(\tilde\theta_\psi) = \partial h(\theta)/\partial\psi|_{\theta=\tilde\theta_\psi}$. An essential step in the derivation of (2.12) is the change of variable from ψ to $r_p^B(\psi)$, which has Jacobian $-h_\psi(\tilde\theta_\psi)/r_p^B(\psi)$ (see Fraser *et al.*, 1999; DiCiccio & Martin, 1991, among others). See also Sweeting (1995, 1996) for alternative derivations of posterior tail area approximations based on the Laplace formula.

Another posterior tail area approximation, equivalent to (2.12), can be obtained by considering expansions around the MLE. In this case, the tail area approximation is obtained by integrating (2.11) for $p = 1$, with the final result (see, *e.g.*, Brazzale *et al.*, 2007, Ch. 8; Reid, 2003)

$$\int_{\psi_0}^{\infty} \pi(\psi|y) d\psi \;=\; \Phi\{r_p^\star(\psi_0)\}\{1 + O(n^{-3/2})\}, \qquad (2.13)$$

where $r_p^\star(\psi) = r_p(\psi) + r_p(\psi)^{-1} \log\{q_B(\psi)/r_p(\psi)\}$ is the modified likelihood root, $r_p(\psi) = \mathrm{sign}(\hat\psi - \psi)[2(\ell_p(\hat\psi) - \ell_p(\psi))]^{1/2}$ is the likelihood root, and

$$q_B(\psi) \;=\; \ell_p'(\psi)\, j_p(\psi)^{-1/2}\, \frac{|j_{\lambda\lambda}(\psi, \hat\lambda_\psi)|^{1/2}}{|j_{\lambda\lambda}(\hat\psi, \hat\lambda)|^{1/2}} \frac{\pi(\hat\psi, \hat\lambda)}{\pi(\psi, \hat\lambda_\psi)}, \qquad (2.14)$$

with $\ell_p'(\psi) = \partial\ell_p(\psi)/\partial\psi$ the profile score.

While the approximation (2.10), its associated tail area approximation (2.12), and the corresponding versions based on the MLE all have the same theoretical approximation error, the versions based on the expansion of $h(\theta)$ tend to be more accurate in practice.

However, the MLE-based versions are easier to implement in practice because they involve only standard likelihood quantities, *e.g.* the MLE, the constrained MLE and the observed information, which are readily available from any software that performs numerical optimizations. When $\pi(\theta) \propto 1$, expressions (2.10) and (2.12) coincide with their respective approximations based on log-likelihood expansions.

When the class of matching priors (Tibshirani, 1989; Datta & Mukerjee, 2004) is considered, *e.g.* prior distributions under which the posterior probabilities of certain regions coincide with their coverage probabilities either exactly or approximately, in (2.4), the marginal posterior distribution for ψ can be expressed as (Ventura *et al.*, 2009, 2013)

$$\pi(\psi|y) \propto L_{mp}(\psi)\pi_{mp}(\psi), \tag{2.15}$$

where $L_{mp}(\psi) = L_p(\psi)M(\psi)$ is the modified profile likelihood for a suitably defined correction term $M(\psi)$ (see, among others, Severini, 2000, Ch. 9 and Pace & Salvan, 2006), and the corresponding matching prior is

$$\pi_{mp}(\psi) \propto i_{\psi\psi.\lambda}(\psi, \hat{\lambda}_\psi)^{1/2}, \tag{2.16}$$

with $i_{\psi\psi.\lambda}(\psi,\lambda) = i_{\psi\psi}(\psi,\lambda) - i_{\psi\lambda}(\psi,\lambda)i_{\lambda\lambda}(\psi,\lambda)^{-1}i_{\psi\lambda}(\psi,\lambda)^T$ being the partial information, and where $i_{\psi\psi}(\psi,\lambda)$, $i_{\psi\lambda}(\psi,\lambda)$, and $i_{\lambda\lambda}(\psi,\lambda)$ are blocks of the expected Fisher information $i(\psi,\lambda)$ from $\ell(\psi,\lambda)$. Starting from (2.16), it is possible to show that (see, *e.g.* Ventura *et al.*, 2013) (2.13) holds with $r_p^\star(\psi)$ given by the modified profile likelihood root of Barndorff-Nielsen & Chamberlin (1994); see also Barndorff-Nielsen & Cox (1994) and Severini (2000, Ch. 7). In particular, the quantity (2.14) has the form

$$q_B(\psi) = \frac{\ell_p'(\psi)}{j_p(\hat{\psi})^{1/2}} \frac{i_{\psi\psi.\lambda}(\hat{\psi},\hat{\lambda})^{1/2}}{i_{\psi\psi.\lambda}(\psi,\hat{\lambda}_\psi)^{1/2}} \frac{1}{M(\psi)} .$$

The posterior tail area approximations (2.12) and (2.13) and the tail area based on the matching posterior (2.15) can be used to compute marginal posterior probabilities or posterior quantiles. For instance, in the case of (2.12), the approximate posterior median is found by solving $r_B^\star(\psi) = 0$, whereas an approximate equi-tailed $(1-\alpha)$ credible interval can be obtained as

$$CI_{1-\alpha} = \{\psi : |r_B^\star(\psi)| \leq z_{1-\alpha/2}\},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of the standard normal distribution.

As shown in formula (2.7), the Laplace approximation of the normalizing constant $p(y)$ has asymptotic error of order $O(n^{-1})$. Higher-order approximations of $p(y)$ can be obtained through Bayesian Bartlett corrections. In particular, DiCiccio *et al.* (1997)

proposed the Bartlett-corrected Laplace approximation

$$p_B(y) = p_L(y) \left[\frac{E_\pi\{w(\theta)\}}{d} \right]^{d/2}, \tag{2.17}$$

with $w(\theta) = 2\{h(\tilde{\theta}) - h(\theta)\}$, where the expectation is typically performed via Monte Carlo methods. This approximation has asymptotic error of order $O(n^{-2})$ (see DiCiccio *et al.*, 1997).

In models with few parameters, *e.g.* $d \leq 5$, numerical integration methods based on quadrature rules can also be used. Briefly, the aim of quadrature rules is to divide the integrand in smaller pieces, compute their areas and approximate the integral by the sum of the areas. Quadrature rules for posterior approximations were first proposed by Naylor & Smith (1982). They are typically very accurate, but unfortunately their use is limited by the curse of dimensionality, since the computational complexity increases rapidly with the number of the parameters (see, *e.g.*, Evans & Swartz, 1995, 2000).

Posterior approximations via the Laplace expansion are generally fast, accurate and easy to implement. However, when routinely applied, these posterior approximations typically require new optimization and (numerical) differentiation tasks. In light of this, posterior approximations via the Laplace expansion can be time consuming and may encounter numerical issues, especially for models with many parameters. These issues can be avoided if one first samples from the approximation of the posterior parametrized on a computationally convenient scale, and numerically transforms the values in the required scale. This idea is developed in Section 3.1, where an algorithm is proposed to quickly simulate from the approximate marginal posterior density (2.10) by using the posterior tail area approximation (2.12).

2.4 Monte Carlo methods

In complicated and highly multidimensional posteriors, quadrature and asymptotic methods, such as those presented so far may be not applicable, *e.g.* the regularity conditions may not hold. In these cases, Monte Carlo or stochastic methods are the only alternative (Evans & Swartz, 1995). The aim of Monte Carlo methods is to approximate difficult integrals via stochastic simulation from appropriate probability distributions and ultimately to approximate integrals with finite sums.

A general advantage of Monte Carlo methods is that the associated approximations are typically simulation consistent, *e.g.* for a large number of simulations the approximations will converge to the true value. However, the approximation error is not always easy to control, especially in Monte Carlo methods which produce dependent samples.

The basic Monte Carlo approximation for integrals works as follows. Suppose it is of interest to compute the integral $E_f(g) = \int g(x)f(x)\,dx$, where $g(x)$ is a given function and $f(x)$ a proper probability distribution. Then, given a sample $(x^{(1)}, \ldots, x^{(m)})$ drawn from $f(x)$, the Monte Carlo approximation of $E_f(g)$ is

$$\bar{g}_m = \frac{1}{m}\sum_{t=1}^{m} g(x^{(t)}).$$

By the Strong Law of Large Numbers (SLLN), $\bar{g}_m$ converges almost surely to $E_f(g)$. Moreover, if the square of $g(x)$ has finite expectation, *i.e.* $E_f(g^2) < \infty$, the speed of convergence can be assessed via the Central Limit Theorem (see, *e.g.*, Robert & Casella, 2004, Ch. 3).

2.4.1 Monte Carlo methods for posterior computation

At heart of Monte Carlo methods is the ability to simulate from probability distributions or in Bayesian terms, the ability to simulate from the posterior. Indeed, provided a sample from the posterior distribution is available, we can, in principle, estimate any posterior summary. However, in practice, sampling from $\pi(\theta|y)$ can be cumbersome.

There exists a wide variety of posterior simulation techniques, among which Markov chain Monte Carlo (MCMC) or importance sampling (IS) methods are the most widely used (see, *e.g.*, Evans & Swartz, 1995, 2000; Robert & Casella, 2004).

IS is perhaps the easiest method, at least for posteriors with moderate numbers of parameters. Consider the posterior expectation (2.3), and suppose its computation is impossible because of the unknown normalizing constant $p(y)$. Let $f(\theta)$ be a probability density, *i.e.*, the importance density, which is straightforward to sample from and is such that its support includes that of $\pi(\theta|y)$. Then, given a sample $(\theta^{(1)}, \ldots, \theta^{(m)})$ from $f(\theta)$, the IS approximation of (2.3) is

$$\bar{g}_m = \frac{\sum_{t=1}^{m} g(\theta^{(t)})w(\theta^{(t)})}{\sum_{t=1}^{m} w(\theta^{(t)})},$$

where $w(\theta) = \pi(\theta|y)/f(\theta)$ are the importance weights. By the SLLN $\bar{g}_m$ converges to $E_\pi(g)$ almost surely. Provided $E_\pi(g^2) < \infty$, the speed of convergence of $\bar{g}_m$ can be assessed via the Central Limit Theorem (see, *e.g.*, Evans & Swartz, 2000, p. 173).

The IS method is the basis of the *sampling/importance resampling* (SIR) algorithm, which can be used to simulate approximate samples from $\pi(\theta|y)$. This proceeds by generating samples $(\theta^{(1)}, \ldots, \theta^{(m)})$ from $f(\theta)$ and then resampling, with replacement from this sample with weights $w_t \propto \pi(\theta^{(t)}|y)/f(\theta^{(t)})$, $t = 1, \ldots, m$, a new sample $(\theta^{(1*)}, \ldots, \theta^{(m*)})$,

which can be treated as a sample from $\pi(\theta|y)$. However, the posterior summaries computed from the sample generated with the SIR algorithm may have higher variance than the ones obtained with IS (Evans & Swartz, 2000, p. 175).

The performance of the IS and the SIR algorithms depends on the importance density $f(\theta)$, whose tails should be heavier than those of $\pi(\theta|y)$. Of course, this is very difficult to check in practice. Typical importance densities are multivariate normals or multivariate t-student distributions. Variance reduction techniques can be used in order to decrease the variability of IS. Finally, we notice that IS can be embedded in an iterative procedure where the importance function, taken from a certain family of distributions, is progressively improved until some criterion is reached. This procedure, called Adaptive Importance Sampling, in general improves upon the usual IS; see Evans & Swartz (2000, Chap. 6) for more details.

Another popular way of simulating from a general posterior distribution is by using MCMC methods. The MCMC sampling strategy sets up an irreducible, aperiodic Markov chain whose stationary distribution equals the posterior distribution of interest. A general way of constructing a Markov chain is by using a Metropolis-Hastings algorithm. Here, we focus on two particular variants of Metropolis-Hastings algorithms, the independence chain and the random walk chain, that are applicable to a wide variety of Bayesian inference problems.

Suppose it is of interest to simulate from a posterior density $\pi(\theta|y)$. A Metropolis-Hastings algorithm begins with an initial value $\theta^{(0)}$ and specifies a rule for simulating the tth value in the sequence $\theta^{(t)}$, given the $(t-1)$th value in the sequence $\theta^{(t-1)}$. This rule consists of a proposal density $q(\cdot|\cdot)$, which simulates a candidate value θ^*, and of the computation of an acceptance probability P, which indicates the probability that the candidate value will be accepted as the next value in the sequence. Specifically, this

algorithm can be described as follows:

Data: Starting values $\theta^{(0)}$, proposal distribution $q(\cdot|\cdot)$
Result: A dependent sample $(\theta^{(1)}, \ldots, \theta^{(m)})$ from $\pi(\theta|y)$
for $t = 1 \to m$ **do**
  draw $\theta^* \sim q(\theta|\theta^{(t-1)})$ and compute

$$R_t = \frac{\pi(\theta^*)L(\theta^*)q(\theta^{(t-1)}|\theta^*)}{\pi(\theta^{(t-1)})L(\theta^{(t-1)})q(\theta^*|\theta^{(t-1)})}$$

  compute $P_t = \min\{R_t, 1\}$.
  Draw $W_t \sim \text{Bernoulli}(P_t)$
  if $W_t = 1$ **then**
    acccept θ^* and set $\theta^{(t)} = \theta^*$
  else
    set $\theta^{(t)} = \theta^{(t-1)}$
  end
end

Algorithm 1: Metropolis-Hastings

Under some regularity conditions on the proposal $q(\cdot|\cdot)$, on the prior and on the likelihood, the sequence of simulated values $(\theta^{(1)}, \ldots, \theta^{(m)})$ will converge to a random variable that is distributed according to the posterior distribution $\pi(\theta|y)$; see, for instance, Robert & Casella (2004, Ch. 7).

Different Metropolis-Hastings algorithms are constructed depending on the choice of proposal density. For instance, if the proposal density is independent of the current value in the sequence, that is if

$$q(\theta^*|\theta^{(t-1)}) = q(\theta^*),$$

then the resulting algorithm is called an independence chain. Other proposal densities can be defined by letting the density have the form

$$q(\theta^*|\theta^{(t-1)}) = s(\theta^* - \theta^{(t-1)}),$$

where $s(\cdot)$ is a symmetric density about the origin. This type of chain is a random walk chain and the ratio R_t has the simple form

$$R_t = \frac{\pi(\theta^*)L(\theta^*)}{\pi(\theta^{(t-1)})L(\theta^{(t-1)})}.$$

Which features the proposal density must have depends on the type of MCMC algorithm employed. For an independence chain, we desire that the proposal density $q(\theta)$ approximates the posterior density $\pi(\theta|y)$, suggesting a high acceptance rate. But, the ratio $\pi(\theta|y)/q(\theta)$ must be bounded, especially in the tail portion of the posterior density. For random walk chains with normal proposal densities, it has been suggested that acceptance rates between 25% and 45% are good.

When the posterior $\pi(\theta|y)$ admits conditional distributions which are known densities, apart from the normalizing constants, then Gibbs sampling is an easier alternative to Metropolis-Hastings algorithms. Let us define the set of full conditional densities as

$$\pi(\theta_i|\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d, y), \quad i = 1, \ldots, d, \tag{2.18}$$

where θ_i denotes the ith element of the parameter θ. It is assumed that the full conditionals are proper and easy to sample from. The idea behind Gibbs sampling is that we can set up a Markov chain simulation algorithm from the joint posterior distribution by successfully simulating individual parameters from the set of d full conditional distributions, as follows:

Data: Starting values $\theta^{(0)}$ and the full conditional distributions
Result: A dependent sample $(\theta^{(1)}, \ldots, \theta^{(m)})$ from $\pi(\theta|y)$
for $t = 1 \to m$ **do**
1 draw $\theta_1^{(t)} \sim \pi(\cdot|\theta_2^{(t-1)}, \ldots, \theta_d^{(t-1)}, y)$
2 draw $\theta_2^{(t)} \sim \pi(\cdot|\theta_1^{(t)}, \theta_3^{(t-1)} \ldots, \theta_d^{(t-1)}, y)$
⋮ ...
d draw $\theta_d^{(t)} \sim \pi(\cdot|\theta_1^{(t)}, \ldots, \theta_{d-1}^{(t)}, y)$
end

Algorithm 2: Gibbs sampling

Simulating one value of each individual parameter from the full conditionals (2.18) in turn is called one cycle of Gibbs sampling. Under general conditions, draws from this simulation algorithm will converge to the target distribution $\pi(\theta|y)$. In situations where it is not convenient to sample directly from the conditional distributions, one can use a Metropolis-Hastings algorithm, such as the random walk, to simulate from each distribution.

Once a posterior sample is available, the marginal posterior of a parameter of interest ψ can be easily computed with kernel density estimation methods (see, *e.g.* Silverman, 1986), or via the *Rao-Blackwellization* method of Gelfand & Smith (1990), when the conditional distributions are known. See also Chen (1994) for a generalization of the Gelfand & Smith's method.

2.4.2 Monte Carlo methods for marginal likelihoods

There are many approaches to the computation of posterior model probabilities (see, *e.g.*, Chen *et al.*, 2000, Ch. 5). Here we are mostly concerned with the calculation of posterior normalizing constants, which are the ingredients for computing BFs as well as posterior model probabilities.

In principle, the Monte Carlo method can be used to approximate also the normalizing constant (2.2) via the empirical mean

$$\bar{p}(y) \;=\; \frac{1}{m}\sum_{t=1}^{m} L(\theta^{(t)}), \tag{2.19}$$

where $(\theta^{(1)}, \ldots, \theta^{(m)})$ is a sample from $\pi(\theta)$. Although this method is easy to implement, in practice it requires an enormous number of simulations in order to obtain accurate estimates (Lewis & Raftery, 1997), with the result of preventing its use in routine analyses.

A first estimator for (2.2), which uses simulation from the posterior, is the harmonic mean estimator (Newton & Raftery, 1994)

$$\bar{p}_{NR}(y) \;=\; \left\{ \frac{1}{m}\sum_{t=1}^{m} \frac{1}{L(\theta^{(t)})} \right\}^{-1}.$$

Unfortunately, this estimator is not stable, since the inverse of the likelihood does not have finite variance (Chib, 1995). Gelfand & Dey (1994) propose a generalization of the harmonic mean estimator, given by

$$\bar{p}_{GD}(y) \;=\; \left\{ \frac{1}{m}\sum_{t=1}^{m} \frac{g(\theta^{(t)})}{L(\theta^{(t)})\pi(\theta^{(t)})} \right\}^{-1}, \tag{2.20}$$

where $g(\theta)$ is a density with tails thinner than the kernel $\pi(\theta)L(\theta)$ of the posterior. Clearly, in the particular case with $g(\theta) = \pi(\theta)$, the estimator (2.20) coincides with the harmonic mean estimator. Although (2.20) solves the problem of instability of the harmonic mean estimator, it requires a tuning function $g(\theta)$, which can be difficult to determine and to monitor in high-dimensional problems (Chib, 1995).

The normalizing constant can be estimated also with the usual IS method. Given an importance density $f(\theta)$ and a sample of m values from it, the importance sampling estimator of $p(y)$ is

$$\bar{p}_{IS}(y) = \sum_{t=1}^{m} \frac{w(\theta^{(t)})L(\theta^{(t)})}{\sum_{t=1}^{m} w(\theta^{(t)})}, \tag{2.21}$$

where $w(\theta) = \pi(\theta)/f(\theta)$. The performance of the importance sampling estimator depends on the importance density. Poor choices of importance densities may give very misleading results (see, Evans & Swartz, 2000, Ch. 6).

A popular method for computing posterior normalizing constants was proposed by Chib (1995) and successively extended by Chib & Jeliazkov (2001). The method begins by simply rewriting Bayes' rule as

$$p(y) = \frac{L(\theta)\pi(\theta)}{\pi(\theta|y)}. \tag{2.22}$$

Only the denominator or the right-hand side of (2.22) is unknown, so an estimate of the posterior would produce an estimate of the normalizing constant. But since the identity (2.22) holds for any θ, we require only a posterior density estimate at a single point θ^*. So we have (Chib, 1995)

$$\log \hat{p}(y) = \ell(\theta^*) + \log \pi(\theta^*) - \log \hat{\pi}(\theta^*|y), \tag{2.23}$$

where the log scale is convenient for computational accuracy. Although θ^* may be any point of the parametric space, Chib (1995) suggests choosing it as a point of high posterior density, to maximise accuracy in (2.23).

To show how to obtain $\hat{\pi}(\theta^*|y)$, suppose the complete parameter can be decomposed in two blocks $\theta = (\theta_1, \theta_2)$, where $\pi(\theta_2|\theta_1, y)$ and $\pi(\theta_1|\theta_2, y)$ are both available in closed form. In writing

$$\pi(\theta_1^*, \theta_2^*|y) = \pi(\theta_2^*|\theta_1^*, y,)\pi(\theta_1^*|y), \tag{2.24}$$

we observe that the first term on the right-hand side is available explicitly at $\theta^* = (\theta_1^*, \theta_2^*)$, while the second can be estimated via the *Rao-Blackwellization* method, namely

$$\hat{\pi}(\theta_1^*|y) = \frac{1}{m}\sum_{i=1}^{m} \pi(\theta_1^*|\theta_2^{(i)}, y). \tag{2.25}$$

Thus the estimated posterior normalizing constant in (2.23) becomes (see Chib, 1995)

$$\log \hat{p}_C(y) \;=\; \ell(\theta_1^*, \theta_2^*) + \log \pi(\theta_1^*, \theta_2^*) - \log \pi(\theta_2^*|\theta_1^*, y) - \log \hat{\pi}(\theta_1^*|y).$$

The extension to more than two parameter blocks requires additional sampling, besides sampling from the full conditionals. Chib's method is generally applicable when Gibbs' sampling can be used (Chib, 1995), since in this case we can easily simulate from the conditionals. A more general version, in which the sampling from intractable conditionals is done via Metropolis-Hastings algorithms, is presented in Chib & Jeliazkov (2001).

Many other simulation-based methods for computing marginal likelihoods can be found

in the literature (see Friel & Wyse, 2012, for an extensive review), since the topic is currently an active area of research.

In Section 3.4 we present an original contribution to the computation of posterior normalizing constants based on the combination of the Laplace approximation for marginal posterior distributions (Tierney & Kadane, 1986) with Chib's idea. The proposed method has relative error of order $O(n^{-3/2})$ and does not require posterior simulation.

2.5 Methods for complex models

So far it was assumed that the likelihood $L(\theta)$ is analytically or computationally tractable. However, various modern applications which involve models with complex dependence structures, models with many latent variables or semi-parametric models, likelihood-based methods may encounter computational problems, due to the difficulty of evaluating $L(\theta)$. This difficulty poses a serious obstacle to all likelihood-based inference methods, and all the techniques presented so far are of little or no use.

Here we summarize two approaches useful for dealing with complex models. The first is based on the use of approximate likelihoods directly in the Bayes' rule, as if they were proper likelihood functions, *i.e.*, likelihoods derived from the density of the data. The second approach uses the so called likelihood-free or Approximate Bayesian Computation (ABC) methods, which try to approximate the likelihood, and hence the posterior, by simulating pseudo-datasets from the model.

2.5.1 Pseudo-likelihood methods

In the Bayesian framework, complex models can be usefully handled by using a posterior distribution based on the combination of a suitable pseudo-likelihood function with a prior distribution, as indicated by the growing interest in the statistical literature.

A general pseudo-likelihood $\tilde{L}(\theta) = \tilde{L}(\theta; y)$ based on the data $y = (y_1, \ldots, y_n)$ is a function of the parameter θ, with properties similar to those of a genuine likelihood function. For instance, the pseudo-score function has zero null expectation, and the maximum pseudo-likelihood estimator (MPLE) is consistent and asymptotically normally distributed (see Pace & Salvan, 1997, Ch. 4 and Severini, 2000, Ch. 8). By considering a pseudo-likelihood $\tilde{L}(\theta)$ and a prior $\pi(\theta)$, a pseudo-posterior distribution can be defined as

$$\tilde{\pi}(\theta|y) \propto \tilde{L}(\theta)\pi(\theta). \tag{2.26}$$

Pseudo-posterior distributions of the form (2.26) have been discussed, for instance, in the Bayesian literature for the elimination of nuisance parameters (see Chang & Mukerjee, 2006; Ventura *et al.*, 2009; Chang *et al.*, 2009; Racugno *et al.*, 2010; Ventura *et al.*, 2013, among others). In this context the pseudo-likelihood approach has the advantage of not requiring elicitation on the nuisance parameters, which in general may be difficult. Since the posterior distribution, and the BFs, may be sensitive to the chosen prior, clearly a pseudo-likelihood approach which requires priors only on the parameter of interest is extremely useful. Other uses of pseudo-likelihoods for posterior inference refer to robustness with respect to the presence of outliers or model misspecification (Greco *et al.*, 2008; Ventura *et al.*, 2010; Agostinelli & Greco, 2013) or to relieve some assumptions on the model (Raftery *et al.*, 1996; Lazar, 2003; Lin, 2006; Pauli *et al.*, 2011; Ribatet *et al.*, 2012; Yin, 2009, and references therein).

An example of approximate likelihoods useful for dealing with complex models is the class of *composite likelihoods*, which are based on the composition of suitable lower dimensional densities, such as bivariate, conditional or full conditional densities (Varin, 2008; Varin *et al.*, 2011), or even a combination of them. In particular, let $y = (y_1, \ldots, y_n)$ be a random sample from $Y_i \sim p(y_i; \theta)$, 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 $L_k(\theta; y_i) \propto p(y \in A_k(y_i); \theta)$ can be computed. The composite log-likelihood is defined as

$$c\ell(\theta; y) = \sum_{i=1}^{n} \sum_{k=1}^{K} w_k \log L_k(\theta; y_i), \tag{2.27}$$

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 $p_{hk}(\cdot, \cdot; \theta)$, (2.27) is called a pairwise log-likelihood and is given by

$$p\ell(\theta; y) = \sum_{i=1}^{n} \sum_{\substack{h,k=1 \\ h \neq k}}^{q} w_{hk} \log p_{hk}(y_{ih}, y_{ik}; \theta). \tag{2.28}$$

The validity of inference about θ 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) $\hat{\theta}_c$ is the solution of the composite score equation

$$c\ell_\theta(\theta; y) = \frac{\partial c\ell(\theta; y)}{\partial \theta} = 0. \tag{2.29}$$

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, $\hat{\theta}_c$ is consistent and approximately normal, with mean θ and variance

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

where $K(\theta) = E_\theta\{-\partial c\ell_\theta(\theta; y)/\partial\theta^T\}$ and $J(\theta) = \mathrm{var}_\theta\{c\ell_\theta(\theta; y)\}$ are the sensitivity and the variability matrices, respectively. The matrix $G(\theta)^{-1}$ is known as the Godambe information, and the form of $G(\theta)$ is due to the failure of the information identity since, in general, $K(\theta) \neq J(\theta)$. This failure also implies that the composite likelihood is wrongly too concentrated (see, *e.g.* Pauli *et al.*, 2011).

The asymptotic distribution of the composite log-likelihood ratio $cw(\theta) = 2\{c\ell(\hat{\theta}_c; y) - c\ell(\theta; y)\}$ is a linear combination of independent chi-square random variables, i.e. $cw(\theta) \xrightarrow{d} \sum_{j=1}^{d} \omega_j Z_j^2$, 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 $K(\theta)^{-1} J(\theta)$. In the special case of $d = 1$, we have $\omega_1 = J(\theta)/K(\theta)$, and the adjusted pairwise log likelihood ratio statistic $cw_1(\theta) = cw(\theta)/\omega_1$ is asymptotically χ_1^2. For $d > 1$, first-order moment matching can be used, which gives

$$cw_1(\theta) = \frac{cw(\theta)}{\bar{\omega}} \ , \tag{2.30}$$

with $\bar{\omega} = \sum_{j=1}^{d} \omega_j/d$. A χ_d^2 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).

Pauli *et al.* (2011) suggest to combine the composite likelihood $cL(\theta) = \exp\{c\ell(\theta)\}$ suitably calibrated, i.e.

$$cL_c(\theta) = cL(\theta)^{1/\bar{\omega}}, \tag{2.31}$$

with a prior $\pi(\theta)$ in the Bayesian framework to obtain the calibrated composite posterior distribution

$$\pi_c(\theta|y) \propto \pi(\theta) cL_c(\theta) \ . \tag{2.32}$$

The calibration in (2.31) is necessary in order to adjust the curvature of the composite likelihood (see also Smith & Stephenson, 2009) and allows one to approximately recover the asymptotic properties of the full posterior. Examples of (2.32) are discussed in Pauli *et al.* (2011); see also Ribatet *et al.* (2012). A limitation of (2.32) is that it depends crucially on the calibration factor, whose components are typically cumbersome to compute (see Varin *et al.*, 2011, Section 5.1).

The class of composite likelihoods contains, and thus generalizes, the ordinary likelihood, as well as many other alternatives, such as the pseudo-likelihood of Besag (1974) and Cox's partial likelihood (Cox, 1975).

The partial likelihood was introduced by Cox (1975) as an inferential tool in proportional hazard models (Cox, 1972) with censored observations. The proportional hazard model is widely used for semi-parametric survival data modelling. In its simplest, form the failure times $T_1, \ldots, T_n$, for n independent individuals, have hazard function $h(t; x_i) = h_0(t) \exp\{x_i^T \beta\}$, where $\beta = (\beta_1, \ldots, \beta_d)$ is the regression parameter, x_i is a $(d \times 1)$ vector of covariates for unit i, $(i = 1 \ldots, n)$, and $h_0(t)$ is the baseline hazard function. Suppose that the data are n pairs (t_i, δ_i), where t_i denotes the observed lifetimes for the ith individual and δ_i is an indicator taking value 1 if t_i is uncensored and 0 otherwise $(i = 1, \ldots, n)$. The partial likelihood for β is

$$L_{PA}(\beta) = \prod_{i=1}^{c} \frac{e^{x_i^T \beta}}{\sum_{j \in \mathcal{R}(t_{(i)})} e^{x_j^T \beta}}, \tag{2.33}$$

where $t_{(i)}$ is the ordered failure time, $\mathcal{R}(t_{(i)})$ is the set of the indexes of the individuals at risk in the instant $t_{(i)}$, that is $\mathcal{R}(t_{(i)}) = \{(i), (i+1) \ldots, (n)\}$, and $c = \sum_i \delta_i$ $(i = 1, \ldots, n)$.

Under the Bayesian paradigm, given a prior density $\pi(\beta)$ on the regression parameters and the partial likelihood (2.33), we can derive the pseudo-posterior distribution

$$\pi_{PA}(\beta|y) \propto \pi(\beta) L_{PA}(\beta); \tag{2.34}$$

see Raftery *et al.* (1996), Volinsky *et al.* (1997), Ibrahim *et al.* (2001), and references therein, for various applications of (2.34). A Bayesian justification of (2.34) is due to Kalbfleisch (1978); see also Sinha *et al.* (2003), and Kim & Kim (2009).

Standard Monte Carlo methods, such as those described in Section 2.3, can be used in order to approximate pseudo-posterior distributions. Higher-order asymptotic approximation methods can also be applied to form (2.27), (2.33) and (2.35), provided the regularity conditions are satisfied. However, while there are some examples of the Laplace approximation applied to pseudo-posterior distributions (see, *e.g.*, Pauli *et al.*, 2011), the application of the tail area approximation is unexplored in the Bayesian literature. An application of the tail area to the context of pseudo-posteriors is discussed in Section 3.2.

2.5.2 Likelihood-free methods

Often the simulation from complex models is easy but calculating the full likelihood $L(\theta)$, even by using computationally intensive methods, is impractical. An alternative approach to inference is based on simulations from $p(y; \theta)$ for different parameter values, and on the comparison of simulated datasets with the observed data. The idea is to estimate $L(\theta)$ at a given parameter value from the portion of datasets, simulated at that

parameter value, that are similar to the observed data. This is an old idea which was first advocated by Diggle & Gratton (1984).

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). The primary purpose of ABC algorithms is to approximate the posterior distribution, when usual methods, such as MCMC, Gibbs sampling, IS or Laplace approximation, cannot be used, but when the datasets can be easily simulated at specific parameter values.

The original accept-reject ABC algorithm works by first drawing a candidate parameter value θ^* from the prior. Then a new dataset y is drawn from the model at θ^*. Finally, if the simulated data y are equal to observed y^{obs}, θ^* is accepted. With continuous data the equality among y and y^{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 3), 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 ϵ a tolerance threshold.

Result: A sample $(\theta^{(1)}, \ldots, \theta^{(m)})$ from $\pi(\theta|\eta(y^{\text{obs}}))$
for $i = 1 \rightarrow m$ **do**
  repeat
1 draw $\theta^* \sim \pi(\theta)$
2 draw $y \sim p(y; \theta^*)$
  until $\rho\{\eta(y), \eta(y^{obs})\} \leq \epsilon$;
3 set $\theta^{(i)} = \theta^*$
end

Algorithm 3: ABC accept-reject sampler.

Algorithm 3 samples from the marginal in y of the joint distribution

$$\pi_\epsilon(\theta, y|\eta(y^{\text{obs}})) = \frac{\pi(\theta)p(y;\theta)\mathbb{I}_{A_{\epsilon,y^{\text{obs}}}}(y)}{\int_{A_{\epsilon,y^{\text{obs}}}\times\Theta} \pi(\theta)p(y;\theta)\,dyd\theta}, \tag{2.35}$$

where $\mathbb{I}_{A_{\epsilon,y^{\text{obs}}}}(y)$ 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: ϵ, $\eta(\cdot)$, and the Monte Carlo error.

The threshold ϵ cannot be fixed to zero, as in this case the probability of accepting a value is also zero. Instead, ϵ is generally set to the αth quantile of the distance among the statistics, with α 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, because simulations from $\pi(\theta)$ do not account for the data at the proposal stage, and thus 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), IS (Fearnhead & Prangle, 2012), sequential or population Monte Carlo approaches (Sisson *et al.*, 2007, 2009; Beaumont *et al.*, 2009). Hence, the most crucial point of the ABC algorithm is the choice of $\eta(\cdot)$. Indeed, what ABC can achieve at best is $\pi(\theta|\eta(y^{\text{obs}}))$, since $\eta(\cdot)$ is rarely sufficient. This loss of information seems to be a necessary price to pay for the access to computable quantities.

We illustrate two original contributions to the likelihood-free approach in Chapter 4. The first is a contribution in the choice of a default summary statistic, through score and composite score functions. The second contribution concerns the construction of a default proposal distribution for MCMC or IS-type algorithms for ABC.

Chapter 3

Contributions on Asymptotic Posterior Approximations

The approximation of the posterior distribution is one of the fundamental difficulties with the Bayesian approach. For such an approximation MCMC methods are typically used (see, *e.g.*, Robert & Casella, 2004). However, MCMC methods in practice may need to be specifically tailored to the particular model (*e.g.* choice of proposal, convergence checks, etc.) and they may have poor tail behaviour, especially when the dimension of the parameter d is large. Parallel with these simulation-based procedures has been the development of analytical higher-order approximations for parametric inference in small samples (see, *e.g.*, Brazzale & Davison, 2008, and references therein). Using higher-order asymptotics it is possible to avoid difficulties related to MCMC methods and to obtain accurate approximations to posterior distributions, and to the related tail area probabilities (see, *e.g.*, Reid, 1996, 2003; Sweeting, 1996; Brazzale *et al.*, 2007). These methods are highly accurate in many situations, but are nevertheless underused compared to simulation-based procedures (Brazzale & Davison, 2008).

In this chapter we present four developments on higher-order asympotics for Bayesian computations. In particular, Section 3.1 presents the Higher-Order Tail Area (HOTA) sampling scheme, which is useful for simulating values from the approximate posterior distribution of a scalar parameter of interest. Section 3.2 develops an higher-order tail area approximation for pseudo-posterior distributions for a scalar parameter of interest, also with the HOTA sampling scheme. Section 3.3 presents an asymptotic expansion for computing accurate Bayesian credible sets, which have also good sampling properties. Finally, Section 3.4 presents an improved Laplace approximation for computing posterior normalizing constants.

3.1 HOTA sampling scheme

In this section we discuss an original posterior sampling scheme, which is obtained by combining the higher-order tail area approximations (2.12) and (2.13) with the inverse transform sampler (see, *e.g.*, Robert & Casella, 2004, Ch. 2). The proposed method, called HOTA, gives accurate approximations of marginal posterior distributions, and related quantities, also in the presence of multidimensional nuisance parameters.

The HOTA sampling scheme is straightforward to implement, since it is available at little additional computational cost over simple first-order approximations. It is based on an asymptotic expansion of the log-posterior distribution around the posterior mode. In principle, the whole procedure requires as an input only the log-posterior kernel. The method can be applied to a wide variety of statistical models, with the essential requirement of the posterior mode being unique (see Kass *et al.*, 1990, for precise regularity conditions). When the posterior mode is close to the MLE, then an asymptotic expansion around the MLE can be used. The latter approximation allows the use of standard maximum likelihood routines for Bayesian analysis.

The proposed simulation scheme gives independent samples from a third-order approximation (*e.g.*an approximation with error $O(n^{-3/2})$) to the marginal posterior distribution at a negligible computational cost. These are distinct advantages with respect to MCMC methods, which in general are time consuming and provide dependent samples. Nevertheless, MCMC techniques give samples from the full posterior distribution subject only to Monte Carlo error, provided convergence has been reached. On the other hand, HOTA has an easily bounded Monte Carlo error, while it has an asymptotic error for the approximation to the true marginal posterior distribution, which depends on the sample size. This approximation is typically highly accurate even for small n.

One possible use of the HOTA sampling scheme is for quick prior sensitivity analyses (Kass *et al.*, 1989; Reid & Sun, 2010). Indeed, it is possible to easily assess the effect of different priors on marginal posterior distributions, given the same Monte Carlo error. This is not generally true for MCMC or IS methods, which in general have to be tuned for the specific model and prior.

The use of higher-order tail area approximations for posterior simulation is a novel approach in the Bayesian literature. Other attempts to merge asymptotic approximations with Monte Carlo simulations are discussed in Kharroubi & Sweeting (2010) and Guihenneuc-Jouyaux & Rousseau (2005). In particular, Kharroubi & Sweeting (2010) use a multivariate signed root log-likelihood ratio to obtain a suitable importance function for obtaining posterior samples via IS. Guihenneuc-Jouyaux & Rousseau (2005) consider a combination of the Laplace approximation with MCMC in random effects

models where the Laplace method is used to integrate out the random effects, and parameters of interest are approximated via MCMC. However, the HOTA sampling scheme being based on the inversion of the higher-order posterior tail area approximation, avoids the issues related to IS and does not require convergence checks, as with MCMC samples.

3.1.1 HOTA algorithms

The posterior tail area approximations (2.12) and (2.13) give accurate approximations of quantiles of the marginal posterior distribution, but it is not possible to use them to obtain density-based posterior summaries, such as posterior moments or highest posterior density (HPD) regions. One possibility to obtain posterior summaries could be to integrate numerically the Laplace approximation to the marginal density (2.8) or (2.11). However, even though ψ is scalar, numerical integration may become time consuming since a large number of function evaluations is needed to obtain accurate estimates, especially when d is large. In fact, a first numerical integration is needed to compute the normalizing constant and then several numerical integrations are needed for each required posterior summary.

To avoid these drawbacks we introduce the HOTA simulation scheme, which is based on the combination of (2.12) or (2.13), with inverse transform sampling. Its main advantage is that it gives independent samples with negligible computational time. Indeed, its implementation only requires a few function evaluations (*e.g.*, 50), independently of the number of simulations. As happens in every simulation method, the posterior summaries based on the HOTA simulation scheme are subject to Monte Carlo error of order $O_p(m^{-1/2})$, where m is the number of Monte Carlo samples. On the other hand, since the samples are drawn independently, it is easy to control such Monte Carlo error by taking m large enough. Finally, it is important to note that HOTA samples from a third-order approximation of the marginal posterior distribution, whose accuracy depends on the sample size. However, the approximation is typically highly accurate even for small sample sizes.

The HOTA simulation scheme can be implemented in two version, namely HOTA_π and HOTA_ℓ. The former is based on $r_B^\star(\psi)$ and simulates approximate posterior values by inverting (2.12). The latter is based on $r_p^\star(\psi)$ and simulates by inverting (2.13) (see

Algorithm 4 and Algorithm 5)

Result: Independent and approximate sample $(\psi_1, \ldots, \psi_m)$ from $\pi(\psi|y)$

1 Fix m and draw $z = (z_1, \ldots, z_m) \sim N(0, 1)$

2 find $\psi_l = \psi : r_B^\star(\psi) = \min(z)$ and $\psi_u = \psi : r_B^\star(\psi) = \max(z)$

3 set an equispaced grid $\tilde{\psi}_T = (\psi_l, \ldots, \tilde{\psi} - \delta, \tilde{\psi} + \delta, \ldots, \psi_u)$ of length T, and find the corresponding values of $r_B^\star(\psi)$ evaluated at $\tilde{\psi}_T$ denoted by $r_B^\star(\tilde{\psi}_T) = (r_{B,1}^\star, \ldots, r_{B,T}^\star)$

4 interpolate $(r_B^\star(\tilde{\psi}_T), \tilde{\psi}_T)$ by smoothing splines, where $\tilde{\psi}_T$ is the response

for $t = 1 \to m$ **do**

| set ψ_t equal to the predicted value from the spline evaluated at z_t

end

Algorithm 4: The HOTA$_\pi$

Result: Independent and approximate sample $(\psi_1, \ldots, \psi_m)$ from $\pi(\psi|y)$

1 Fix m and draw $z = (z_1, \ldots, z_m) \sim N(0, 1)$

2 find $\psi_l = \psi : r_p^\star(\psi) = \max(z)$ and $\psi_u = \psi : r_p^\star(\psi) = \min(z)$

3 set an equispaced grid $\hat{\psi}_T = (\psi_l, \ldots, \hat{\psi} - \epsilon, \hat{\psi} + \epsilon, \ldots, \psi_u)$ of length T, and find the corresponding values of $r_p^\star(\psi)$ evaluated at $\hat{\psi}_T$ denoted by $r_p^\star(\hat{\psi}_T) = (r_{p,1}^\star, \ldots, r_{p,T}^\star)$

4 interpolate $(r_p^\star(\hat{\psi}_T), \hat{\psi}_T)$ by smoothing splines, where $\hat{\psi}_T$ is the response

for $t = 1 \to m$ **do**

| set ψ_t equal to the predicted value from the spline evaluated at z_t

end

Algorithm 5: The HOTA$_\ell$

The function $r_B^\star(\psi)$ ($r_p^\star(\psi)$) is monotonically increasing (decreasing) in ψ (see, *e.g.*, Brazzale *et al.*, 2007, Ch. 9). For the equispaced grid, moderate values of T are typically sufficient, *e.g.*, 50, and the extremes of the grid can be found numerically (*e.g.* by secant or Brent's method). Notice that $r_B^\star(\psi)$ ($r_p^\star(\psi)$) has a numerical discontinuity at $\tilde{\psi}$ ($\hat{\psi}$), and it may be necessary to exclude values of the grid in a δ-neighbourhood of $\tilde{\psi}$ (ϵ-neighbourhood of $\hat{\psi}$), of the type $\tilde{\psi} \pm \delta$ ($\hat{\psi} \pm \epsilon$). For instance, we can set $\delta = \varepsilon\, \Sigma_{\psi\psi}(\tilde{\theta})^{1/2}$ ($\epsilon = \varepsilon j_p(\hat{\psi})^{-1/2}$), for some small ε, *e.g.*, 0.3. Essentially, fixing the grid in this way, the instabilities of $r_B^\star(\psi)$ ($r_p^\star(\psi)$) are avoided by the numerical interpolation.

Constrained maximization and computation of the required Hessians are generally straightforward to obtain numerically, whenever code for the likelihood or posterior kernel is available. For many statistical models with diffuse priors, built-in `R` functions (see R Core Team 2013) can sometimes be used to obtain full and constrained likelihood maximization as well as the related profile quantities required for HOTA$_\ell$. For instance, the `glm` function in `R` can handle many generalized linear models, and it offers the `offset`

option for constrained estimation. Therefore, if the model in question belongs to the `glm` class, then all the quantities required in HOTA$_\ell$ can be extracted from it.

If the posterior mode and the MLE, computed with usual maximum likelihood routines, are found to be substantially different, HOTA$_\pi$ is a safer choice and its use is recommended. When using HOTAπ, maximum likelihood routines can be used to find appropriate starting values for the posterior optimization. For instance, if the model is in the `glm` class, starting values for the constrained posterior optimization can be obtained from the `glm` command along with the `offset` used to fix the parameter of interest. More generally, starting values for constrained optimization can be obtained by a linear expansion around the maximum (Cox & Wermuth 1990)

$$\lambda^{start} = \hat{\lambda} + j_{\lambda\lambda}(\hat{\psi}, \hat{\lambda})^{-1} j_{\lambda\psi}(\hat{\psi}, \hat{\lambda})(\hat{\psi} - \psi). \tag{3.1}$$

These are the strategies used in the following examples.

Algorithm 4 (5) approximates (2.4) by simulating independently from the higher-order tail area approximation (2.12) ((2.13)). In this respect, it has an obvious advantage over MCMC methods, which usually are more time consuming. Moreover, MCMC methods typically require more attention from the practitioner (*e.g.* choice of the proposal, convergence checks, etc.). A pitfall of HOTA is that its theoretical approximation error (*i.e.* $O(n^{-3/2})$) is bounded by the sample size. Nonetheless, as it will be shown by means of practical examples, HOTA typically gives very accurate approximations, even in small samples.

3.1.2 Examples

The aim of this section is to illustrate the performance of the HOTA method by three examples. In all but the first example, HOTA is compared with the random walk Metropolis. Prior sensitivity analysis is also considered with HOTA and compared also with MCMC. Prior sensitivity with HOTA is based on the same set of independent random variates, thus giving a comparison of different priors, under the same Monte Carlo error.

In general MCMC methods give autocorrelated samples and it is important to check that the chain has converged to its ergodic distribution (see, *e.g.*, Gelman *et al.*, 2003). In the examples, a multivariate normal proposal is used, suitably scaled in order to have an acceptance rate of 30-40%. Chains of simulations are run for a very large number of iterations, are thinned and the initial observations are discarded. In addition, the convergence is checked by the routines of the `coda` package of `R`. In each example,

10^5 final MCMC samples are considered, all with moderate autocorrelation. These MCMC samples will be considered as the gold standard, even though they are only an approximation of the exact posterior distribution.

Algorithm 4 and Algorithm 5 are implemented with the R software, where the spline interpolation is performed with the command `splinfun`, applied to a grid of 50 values evenly spaced with $\varepsilon = 0.3$. A sample of size 10^5 is taken from all the approximate marginal posteriors. Required derivatives are computed numerically. This may be another source of approximation error, difficult to quantify in practice. Nonetheless, we stress that this is an issue for many statistical applications since numerical derivatives are ubiquitous in statistics. Fortunately, there are many routines which provide accurate numerical derivatives; for instance, the `numDeriv R` package (see Gilbert & Varadhan 2012). The `R` code used in the following examples is available at `http://homes.stat.unipd.it/ventura/?page=Software&lang=IT`. An R package for running HOTA in general regular models is under preparation.

Genetic linkage model

The following scalar parameter problem has been studied also in Kharroubi & Sweeting (2010), among others. It concerns a genetic linkage model with n individuals multinomially distributed into four categories with cell probabilities $\{\frac{1}{2}+\frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\}$, with $\theta \in (0,1)$. There are $n = 20$ animals with cell counts $y = (14, 0, 1, 5)$. Under a uniform prior, the posterior of θ is proportional to the likelihood and is given by

$$\pi(\theta|y) \propto (2+\theta)^{14}(1-\theta)\theta^5, \qquad \theta \in (0,1).$$

There are no nuisance parameters and, since $\pi(\theta) \propto 1$, the tail area approximations (2.12) and (2.13) coincide and simplify to

$$\int_{-\infty}^{\theta_0} \pi(\theta|y)d\theta \doteq \Phi\{r^\star(\theta_0)\},$$

where $r^\star(\theta) = r(\theta) + r(\theta)^{-1}\log\{q_B(\theta)/r(\theta)\}$, $q_B(\theta) = -\ell'(\theta)j(\hat\theta)^{-1/2}$, $\ell'(\theta) = d\ell(\theta)/d\theta$, and $r(\theta) = \text{sign}(\theta-\hat\theta)[2(\ell(\hat\theta)-\ell(\theta))]^{1/2}$. In view of this HOTA_ℓ and HOTA_π coincide.

Figure 3.1 shows the posterior distribution computed with HOTA and the exact posterior distribution $\pi(\theta|y)$. The exact posterior distribution appears to be extremely skewed to the right, with a long left tail, and in this case one might expect the HOTA algorithm to fail. On the contrary, it gets very close to the exact posterior, even though the sample

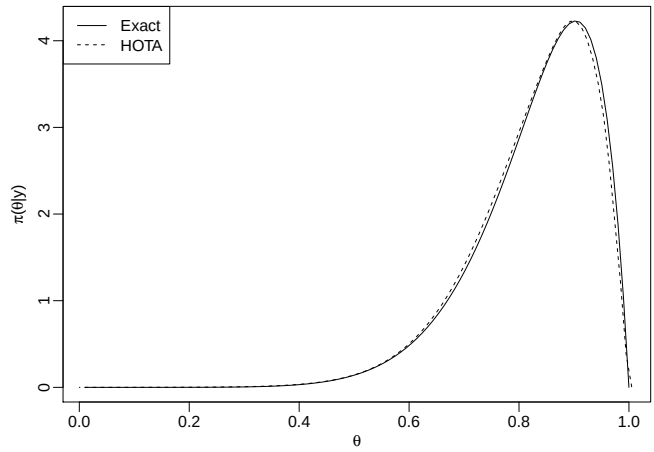


FIGURE 3.1: Genetic linkage model. Exact and HOTA posterior distributions.

size is only $n = 20$. In order to further explore the accuracy of the approximation, the two posteriors are compared also in terms of some summary statistics (mean, standard deviation, 2.5 percentile, median, 97.5 percentile and 0.95 HPD credible set) in Table 3.1. The HOTA results are very close to those based on the exact posterior.

Posterior	Mean	St. Dev.	$Q_{0.025}$	Median	$Q_{0.975}$	0.95 HPD
Exact	0.831	0.108	0.570	0.852	0.978	(0.620, 0.994)
HOTA	0.827	0.108	0.566	0.848	0.976	(0.617, 0.994)

TABLE 3.1: Genetic linkage model. Numerical summaries of the exact and HOTA posterior distributions.

Censored normal regression

The data consist of temperature accelerated life tests on electrical insulation in $n = 40$ motorettes (Davison, 2003, Table 11.10). Ten motorettes were tested at each of four temperatures in degrees Centigrade (150°, 17°, 190° and 220°), the test termination (censoring) time being different at each temperature. These data were analysed from a Bayesian perspective by Kharroubi & Sweeting (2010), among others.

The following linear model is considered

$$y_i = \beta_0 + \beta_1 x_i + \sigma\, \varepsilon_i,$$

where ε_i are independent standard normal random variables, $(i = 1, \ldots, n)$. The response is the $\log_{10}$(failure time), with time in hours, and $x = 1000/(\text{temperature}+273.2)$. Reordering the data so that the first m observations are uncensored, with observed log-failure times y_i, and the remaining $n - m$ are censored at times u_i, the log-likelihood for

$\theta = (\beta_0, \beta_1, \sigma)$ is

$$\ell(\theta) = -m \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - \beta_0 - \beta_1 x_i)^2 + \sum_{i=m+1}^{n} \log \left\{ 1 - \Phi \left(\frac{u_i - \beta_0 - \beta_1 x_i}{\sigma} \right) \right\}. \quad (3.2)$$

For illustrative purposes several prior specifications are considered. The first the flat prior $\pi_F(\theta)$. The second prior is a Normal-Half Cauchy distribution $\pi_{NHC}(\theta)$, given by independent components, which are respectively $\mathrm{N}(0, k)$ for the components of (β_0, β_1) and Half Cauchy with scale s for σ, with $(k, s) = (5, 0.1)$. The third prior is the Zellner's G-prior $\pi_G(\theta)$ (see, *e.g.* Marin & Robert, 2007, Ch. 3), which is the product of σ^{-1} and a bivariate normal density with mean vector a and covariance matrix $c\sigma^2(X^T X)^{-1}$, where X is the design matrix with the first column being a vector of ones. For simplicity we assume $a = (0, 0)$ and $c = 100$. Several proposals exist for fixing c, but we choose 100 since this result can be interpreted as giving to the prior a weight of 1% of the data (see Marin & Robert 2007).

The posterior distributions obtained with these priors do not have a closed form solution, and numerical integration is needed in order to compute $\pi(\psi|y)$, and related quantities, with ψ being a scalar component of θ.

Figure 3.2 shows a sensitivity study on the effect of the three different priors on the posterior distributions based on HOTA$_\pi$. The same set of random variates has been used in all cases, so what is shown are the differences between posteriors, under the same Monte Carlo error. See also Tables 2 and 3 for some numerical summaries for β_1 and σ, respectively. From these illustrations we conclude that the change of the prior influences somehow both σ and β_1.

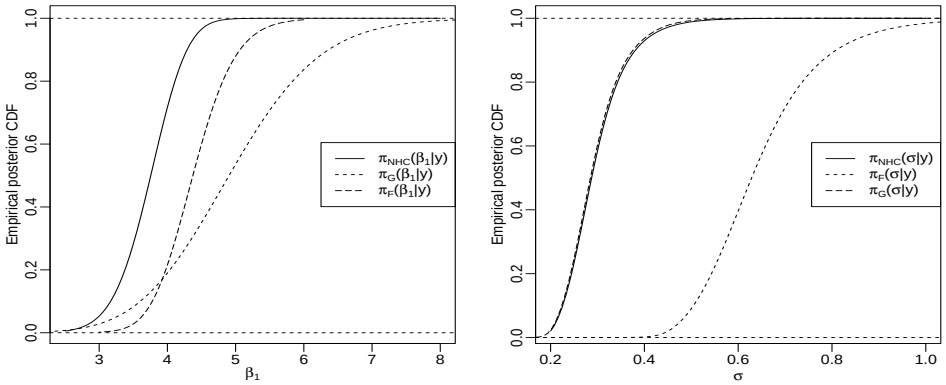


FIGURE 3.2: Censored regression model. Marginal posterior CDFs for β_1 (left) and σ (right), computed with HOTA$_\pi$.

Figure 3.3 presents a graphical comparison between MCMC, HOTA$_\pi$ and HOTA$_\ell$ in terms of the approximate posterior cumulative distribution functions (CDFs) for β_1 (left

column) and σ (right column). Results with HOTA_π are always in close agreement with those of MCMC. On the contrary, the accuracy of HOTA_ℓ may not be satisfactory with non-flat priors, as also confirmed by the summary statistics in Tables 3.2 and 3.3.

Posterior	Method	Mean	St Dev.	$Q_{0.025}$	Median	$Q_{0.975}$	0.95 HPD
$\pi_F(\beta_1\|y)$	MCMC	4.409	0.518	3.461	4.382	5.512	(3.425, 5.470)
	HOTA_ℓ	4.401	0.521	3.459	4.370	5.521	(3.398, 5.443)
	HOTA_π	4.401	0.521	3.459	4.370	5.521	(3.398, 5.443)
$\pi_{NHC}(\beta_1\|y)$	MCMC	3.731	0.447	2.802	3.746	4.571	(2.827, 4.594)
	HOTA_ℓ	3.739	0.437	2.823	3.755	4.549	(2.889, 4.611)
$k = 5$, $s = 0.1$	HOTA_π	3.739	0.443	2.818	3.754	4.569	(2.840, 4.589)
$\pi_G(\beta_1\|y)$	MCMC	4.955	1.114	2.907	4.908	7.304	(2.906, 7.304)
	HOTA_ℓ	5.885	3.078	1.182	5.388	13.173	(0.781, 12.389)
	HOTA_π	4.955	1.099	2.939	4.897	7.285	(2.838, 7.119)

TABLE 3.2: Censored regression model. Numerical summaries of the marginal posteriors of β_1 with $\pi_F(\theta)$, $\pi_{NHC}(\theta)$ and $\pi_G(\theta)$, computed with MCMC, HOTA_ℓ and HOTA_π.

Posterior	Method	Mean	St Dev.	$Q_{0.025}$	Median	$Q_{0.975}$	0.95 HPD
$\pi_F(\sigma\|y)$	MCMC	-1.240	0.201	-1.600	-1.253	-0.811	(-1.616, -0.832)
	HOTA_ℓ	-1.240	0.202	-1.601	-1.251	-0.808	(-1.624, -0.837)
	HOTA_π	-1.240	0.202	-1.601	-1.251	-0.808	(-1.624, -0.837)
$\pi_{NHC}(\sigma\|y)$	MCMC	0.299	0.064	0.201	0.288	0.452	(0.193, 0.431)
	HOTA_ℓ	0.277	0.052	0.196	0.270	0.398	(0.189, 0.384)
$k = 5$, $s = 0.1$	HOTA_π	0.298	0.064	0.203	0.287	0.452	(0.190, 0.426)
$\pi_G(\sigma\|y)$	MCMC	0.649	0.127	0.454	0.630	0.941	(0.434, 0.899)
	HOTA_ℓ	1.327	0.306	0.875	1.278	2.058	(0.815, 1.936)
	HOTA_π	0.647	0.125	0.456	0.628	0.941	(0.430, 0.894)

TABLE 3.3: Censored regression model. Numerical summaries of the marginal posteriors of σ, with $\pi_F(\theta)$, $\pi_{NHC}(\theta)$ and $\pi_G(\theta)$, computed with MCMC, HOTA_ℓ and HOTA_π.

Logistic regression

In this example we consider a logistic regression model applied to the `urine` dataset analysed by Brazzale *et al.* (2007, Ch. 4), among others. This dataset concerns calcium oxalate crystals in samples of urine. The response is an indicator of the presence of such crystals, and the explanatory variables are: specific gravity (`gravity`) (*i.e.* the density of urine relative to water), pH (`ph`), osmolarity (`osmo`, mOsm), conductivity (`conduct`, mMho), urea concentration (`urea`, millimoles per litre), and calcium concentration (`calc`, millimoles per litre). After dropping two incomplete cases, the dataset consists of 77 observations. Let X denote the $(n \times 7)$ design matrix composed by a vector of ones and the six covariates, and let $\beta = (\beta_0, \ldots, \beta_6)$ be regression parameters, where β_0 is the intercept.

Different prior specifications are considered: a flat prior $\pi_F(\beta) \propto 1$, a multivariate normal prior $\pi_N(\beta)$ with independent components $N(a, k)$, with $a = 0$ and $k = 5$, as

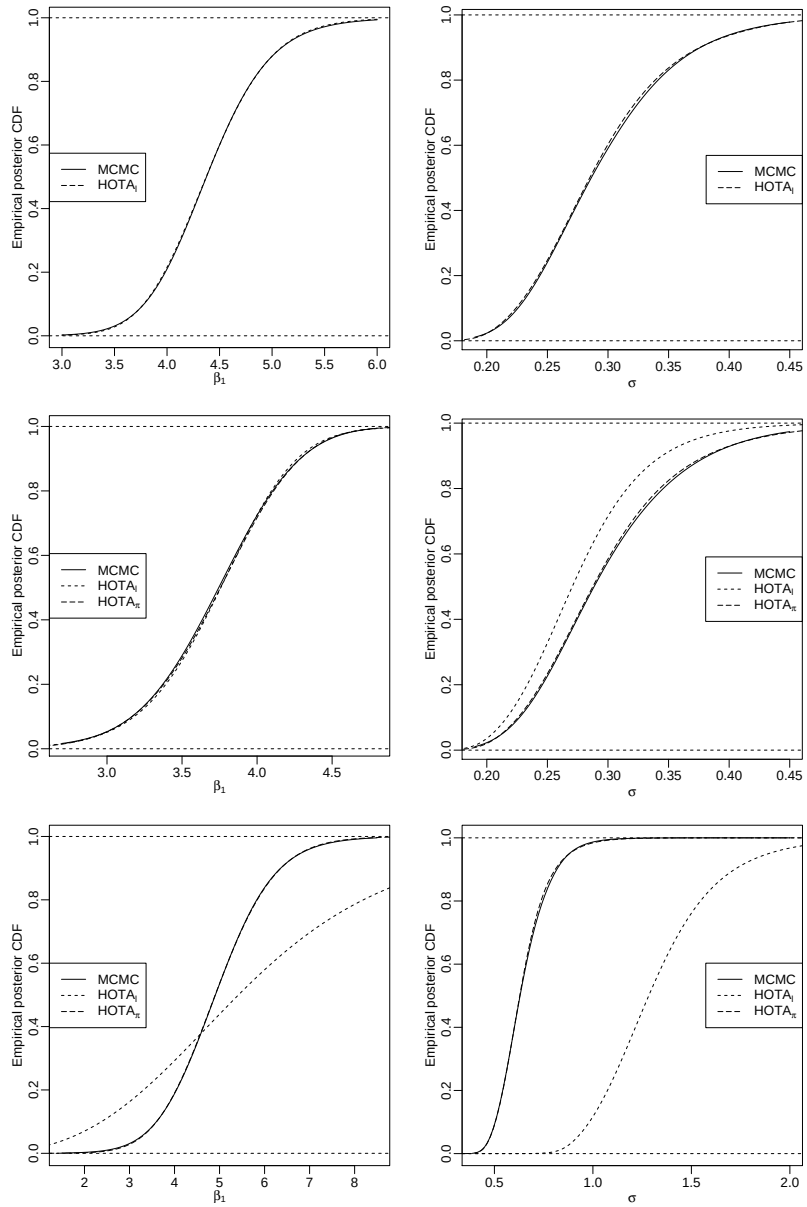


FIGURE 3.3: Censored regression model. Marginal posterior CDFs for β_1 (left column) and σ (right column). The three rows correspond to priors $\pi_F(\theta)$, $\pi_{NHC}(\theta)$ ($k = 5$, $s = 0.1$) and $\pi_G(\theta)$, respectively. In the first line, HOTA$_\pi$ coincides with HOTA$_\ell$.

well as the Zellner's G-prior (see Marin & Robert 2007, Chap. 4), given by

$$\pi_G(\beta) \propto \{\beta^T (X^T X)\beta\}^{-13/4}.$$

The choice of these priors has only the aim of illustrating our method and not to suggest their use for Bayesian data analysis.

Figure 3.4 shows a sensitivity study on the effect of different priors on the posterior distributions based on HOTA$_\pi$. Here, we also consider the matching prior (2.16), given

by

$$\pi_{mp}(\beta_r) \propto j_p(\beta_r)^{1/2}, \quad \text{for } r = 0, \ldots, 6.$$

With this prior the marginal posterior distribution is approximated by HOTA$_\ell$. See also Tables 3.4 and 3.5 for some numerical summaries for β_4 and β_6, respectively.

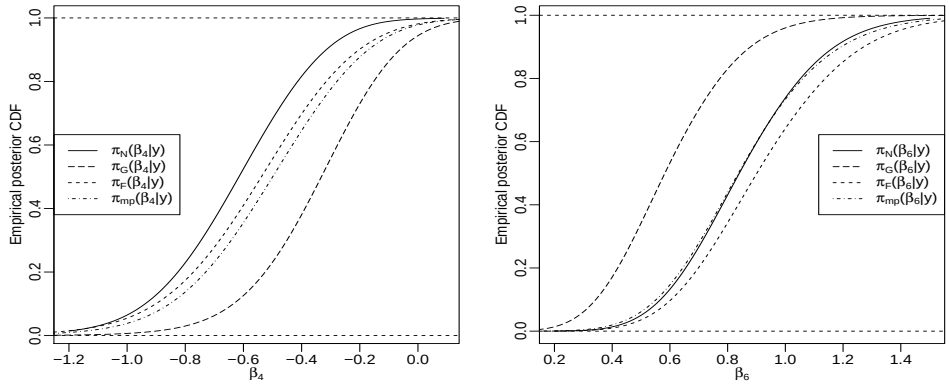


FIGURE 3.4: Logistic regression model. Marginal posterior CDFs for β_4 (left) and β_6 (right), computed with HOTA$_\pi$.

Figure 3.5 presents a graphical comparison between MCMC, HOTA$_\pi$ and HOTA$_\ell$ in terms of posterior cumulative distribution functions (CDF) for β_4 (left column) and β_6 (right column). The same comments about Figure 3.3 apply here, with the difference that the accuracy of HOTA$_\ell$ is better than the previous example when non-flat priors are used. See also Tables 3.4 and 3.5.

Posterior	Method	Mean	St Dev.	$Q_{0.025}$	Median	$Q_{0.975}$	0.95 HPD
$\pi_{mp}(\beta_4\|y)$	HOTA$_\ell$	-0.508	0.270	-1.063	-0.497	-0.007	(-1.010, 0.033)
$\pi_F(\beta_4\|y)$	MCMC	-0.591	0.256	-1.116	-0.585	-0.114	(-1.089, -0.095)
	HOTA$_\ell$	-0.547	0.278	-1.117	-0.537	-0.032	(-1.063, -0.009)
	HOTA$_\pi$	-0.547	0.278	-1.117	-0.537	-0.032	(-1.063, -0.009)
$\pi_N(\beta_4\|y)$	MCMC	-0.619	0.248	-1.132	-0.607	-0.163	(-1.117, -0.155)
	HOTA$_\ell$	-0.645	0.214	-1.073	-0.641	-0.239	(-1.035, -0.206)
$k = 5$	HOTA$_\pi$	-0.623	0.246	-1.127	-0.613	-0.169	(-1.079, -0.133)
$\pi_G(\beta_4\|y)$	MCMC	-0.335	0.227	-0.816	-0.323	0.068	(-0.793, 0.081)
	HOTA$_\ell$	-0.348	0.236	-0.837	-0.336	0.081	(-0.773, 0.114)
	HOTA$_\pi$	-0.343	0.228	-0.819	-0.330	0.070	(-0.790, 0.102)

TABLE 3.4: Logistic regression model. Numerical summaries of the marginal posterior of β_4, with $\pi_{mp}(\beta_4)$, $\pi_F(\beta)$, $\pi_N(\beta)$, and $\pi_G(\beta)$ approximated by MCMC, HOTA$_\ell$ and HOTA$_\pi$.

Posterior	Method	Mean	St Dev.	$Q_{0.025}$	Median	$Q_{0.975}$	0.95 HPD
$\pi_{mp}(\beta_6\|y)$	HOTA_ℓ	0.859	0.255	0.424	0.839	1.417	(0.414, 1.399)
$\pi_F(\beta_6\|y)$	MCMC	0.883	0.250	0.454	0.863	1.425	(0.435, 1.391)
	HOTA_ℓ	0.924	0.264	0.472	0.903	1.500	(0.461, 1.482)
	HOTA_π	0.924	0.264	0.472	0.903	1.500	(0.461, 1.482)
$\pi_N(\beta_6\|y)$	MCMC	0.863	0.241	0.447	0.845	1.386	(0.419, 1.347)
	HOTA_ℓ	0.829	0.217	0.445	0.817	1.289	(0.436, 1.277)
$k = 5$	HOTA_π	0.859	0.239	0.445	0.842	1.373	(0.435, 1.357)
$\pi_G(\beta_6\|y)$	MCMC	0.604	0.204	0.259	0.586	1.054	(0.241, 1.024)
	HOTA_ℓ	0.591	0.197	0.237	0.573	1.030	(0.229, 0.995)
	HOTA_π	0.600	0.212	0.264	0.584	1.060	(0.235, 1.045)

TABLE 3.5: Logistic regression model. Numerical summaries of the marginal posterior of β_6, with $\pi_{mp}(\beta_6)$, $\pi_F(\beta)$, $\pi_N(\beta)$, and $\pi_G(\beta)$ approximated by MCMC, HOTA_ℓ and HOTA_π.

3.1.3 Remarks

The HOTA simulation method for Bayesian approximation combines higher-order tail area approximations with the inverse transform sampler. This sampling method gives accurate approximations of marginal posterior distributions for a scalar parameter of interest.

The accuracy of the two versions of the HOTA algorithm may be different and, in particular, may depend on the chosen prior. In this respect, the version based on the expansion around the posterior mode is a safer choice, since the approximation makes explicit use of the prior information. On the contrary, the accuracy of the version based on the expansion around the MLE, although easier to compute, could be affected by the difference between the likelihood and the posterior, which is indeed the effect of the prior. Therefore, in general we would recommend the use of HOTA_π, since the effect of the prior on the posterior depends on many aspects, such as the nature and range of the parameter, and it is not straightforward to assess such effect in advance. On the other hand, both approximations rely on small-sample results, in the sense that as the sample size increases the effect of the prior vanishes, implying that the two approximations will tend to coincide.

Bayesian robustness with respect to the prior can be easily handled with the HOTA sampling scheme. Indeed, higher-order approximations make it straightforward to assess the influence of the prior, and the effect of changing priors on the posterior quantities (see also Reid & Sun, 2010). Moreover, with HOTA the effect of the prior on the posterior distribution can be appreciated under the same Monte Carlo variation. Finally, default priors, such as the matching prior used in Example 3, could be easily handled by the method and could be used as a benchmark for Bayesian robustness.

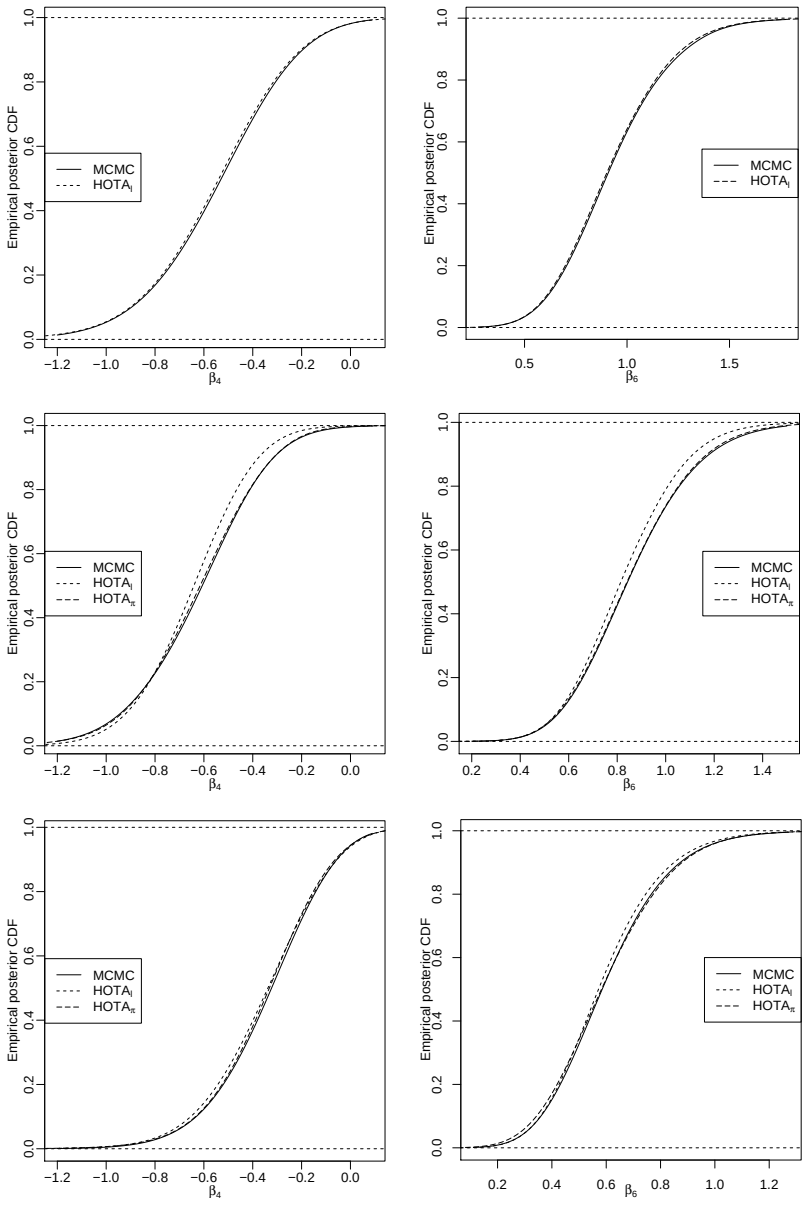


FIGURE 3.5: Logistic regression model. Marginal posterior CDFs for β_4 (left column) and β_6 (right column). The three rows correspond to priors $\pi_F(\beta)$, $\pi_N(\beta)$, with $k = 5$, and $\pi_G(\beta)$ respectively.

The proposed use of higher-order asymptotics for Bayesian simulation opens other interesting applications. For instance, the HOTA procedure could be used in conjunction with MCMC methods, *e.g.*, to simulate from marginal or conditional posteriors within the Gibbs sampling. Moreover, HOTA could be used also to estimate the marginal likelihood (see Sect. 3.4) following the approach of Perrakis *et al.* (2013), where the marginal densities involved can be estimated via kernel methods.

An R package which implements the HOTA algorithm is under preparation, and will soon be available.

3.2 Higher-order tail area approximations for pseudo-posterior distributions

As stated in Section 2.4.1, a possible way to deal with complex models is by means of pseudo-posteriors obtained from the combination of a suitable pseudo-likelihood $\tilde{L}(\theta)$ and a prior for θ within the Bayes' rule.

Let $\theta = (\psi, \lambda)$, where ψ is a scalar parameter of interest and λ is the $(d-1)$-dimensional nuisance parameter (see Racugno *et al.*, 2010, for some examples). In this situation, Bayesian inference about ψ may be based on the marginal pseudo-posterior distribution

$$\tilde{\pi}(\psi|y) = \frac{\int \pi(\psi, \lambda)\, \tilde{L}(\psi, \lambda)\, \mathrm{d}\lambda}{\int\int \pi(\psi, \lambda)\, \tilde{L}(\psi, \lambda)\, \mathrm{d}\lambda\, \mathrm{d}\psi}. \tag{3.3}$$

As for $\pi(\psi|y)$, cumbersome numerical integration may be necessary in order to compute the marginal pseudo-posterior distribution (3.3), in particular when the dimension of λ is large.

This latter difficulty could be avoided using higher-order asymptotics for $\tilde{\pi}(\psi|y)$. In this section, paralleling the results for genuine posterior distributions (see Sec. 2.3.2), a Laplace approximation for (3.3) is presented. Moreover, for a scalar parameter of interest, we derive the corresponding tail area approximation, which can be used to perform accurate Bayesian inference, even for small sample sizes. The methodology proposed can then be used to simulate independent observations from the higher-order approximation of (3.3), paralleling the HOTA algorithms presented in Section 3.1. The advantages of the method are essentially the same as those of HOTA in the context of genuine posterior distributions (see Section 3.1).

3.2.1 Higher-order approximations for $\tilde{\pi}(\psi|y)$

The basic tool for deriving higher-order pseudo-posterior tail area approximations is again the Laplace approximation for integrals (Tierney & Kadane, 1986). Under broad regularity conditions on $\tilde{L}(\theta)$, similar to those required for asymptotic normality of the MLE and under mild regularity conditions on the prior, the Laplace approximation for (3.3) can be obtained in the same way as for genuine posteriors (see Sec. 2.2.2). In particular, assume that $\tilde{L}(\theta) = O(n)$, with a liberal interpretation of n which for independent data is typically given by the sample size, and let $\tilde{h}(\theta) = \log\{\tilde{L}(\theta)\pi(\theta)\}$ be the pseudo log-posterior kernel, which has mode $\tilde{\theta}^\dagger = (\tilde{\psi}^\dagger, \tilde{\lambda}^\dagger)$. Moreover let $\tilde{\theta}^\dagger_\psi = (\psi, \tilde{\lambda}^\dagger_\psi)$ be the constrained mode of $\tilde{h}(\theta)$ with ψ fixed. To approximate the numerator of (3.3),

we expand $\tilde{h}(\psi,\lambda)$ about $\tilde{\lambda}^{\dagger}_{\psi}$ to get

$$\exp\left\{\tilde{h}(\tilde{\theta}^{\dagger}_{\psi})\right\}|\tilde{V}_{\lambda\lambda}(\tilde{\theta}^{\dagger}_{\psi})|^{-1/2},$$

where $\tilde{V}_{\lambda\lambda}(\theta) = -\partial^2\tilde{h}(\psi,\lambda)/\partial\lambda\partial\lambda^T$. Combining this expansion with the Laplace approximation to the denominator, we get

$$\tilde{\pi}(\psi|y) \;=\; \frac{1}{\sqrt{2\pi}}\exp\left\{\tilde{h}(\tilde{\theta}^{\dagger}_{\psi})-\tilde{h}(\tilde{\theta}^{\dagger})\right\}\left\{\frac{|\tilde{V}(\tilde{\theta}^{\dagger})|}{|\tilde{V}_{\lambda\lambda}(\tilde{\theta}^{\dagger}_{\psi})|}\right\}^{1/2}\{1+O(n^{-3/2})\}. \tag{3.4}$$

Note that the approximation error of (3.4) is due to the assumption $\tilde{L}(\theta) = O(n)$.

Paralleling results in Section 2.2.2, for a scalar parameter ψ, formula (3.4) can be integrated to give a tail area approximation. In particular,

$$\begin{aligned}
\int_{-\infty}^{\psi_0}\tilde{\pi}(\psi|y)\mathrm{d}\psi &= \int_{-\infty}^{\psi_0}\frac{1}{\sqrt{2\pi}}\exp\left\{\tilde{h}(\tilde{\theta}^{\dagger}_{\psi})-\tilde{h}(\tilde{\theta}^{\dagger})\right\}\left\{\frac{|\tilde{V}(\tilde{\theta}^{\dagger})|}{|\tilde{V}_{\lambda\lambda}(\tilde{\theta}^{\dagger}_{\psi})|}\right\}^{1/2}\{1+O(n^{-3/2})\}\,d\psi \\
&= \int_{-\infty}^{\tilde{r}_p^B(\psi_0)}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\tilde{r}_p^B(\psi)^2\right\}\left\{\frac{\tilde{r}_p^B(\psi)}{\tilde{q}(\psi)}\right\}\mathrm{d}\tilde{r}_p^B(\psi)\{1+O(n^{-3/2})\} \\
&= \Phi\left\{\tilde{r}_p^B(\psi_0)+\frac{1}{\tilde{r}_p^B(\psi_0)}\log\frac{\tilde{q}_B(\psi_0)}{\tilde{r}_p^B(\psi_0)}\right\}\{1+O(n^{-3/2})\} \\
&= \Phi\left\{\tilde{r}_B^{\star}(\psi_0)\right\}\{1+O(n^{-3/2})\},
\end{aligned} \tag{3.5}$$

where the change of variable from ψ to $\tilde{r}_p^B(\psi) = \mathrm{sign}(\psi-\tilde{\psi}^{\dagger})[2(\tilde{h}(\tilde{\theta}^{\dagger}_{\psi})-\tilde{h}(\tilde{\theta}^{\dagger}))]^{1/2}$ has Jacobian $-\tilde{r}_p^B(\psi)/\tilde{h}_{\psi}(\tilde{\theta}^{\dagger}_{\psi})$ and

$$\tilde{q}_B(\psi) = -\tilde{h}_{\psi}(\psi)\left\{\frac{|\tilde{V}(\tilde{\theta}^{\dagger})|}{|\tilde{V}_{\lambda\lambda}(\tilde{\theta}^{\dagger}_{\psi})|}\right\}^{-1/2}.$$

We notice that an alternative version of the pseudo-posterior tail area approximation can be obtained by expanding the logarithm of the pseudo-likelihood. In this case, the expression for the tail area is similar to (2.13) with all the likelihood-based quantities substituted by the corresponding pseudo-likelihood quantities. In particular, for the Laplace approximation,

$$\tilde{\pi}(\psi|y) \;=\; \frac{|\tilde{j}_p(\hat{\psi})|^{1/2}}{(2\pi)^{p/2}}\exp\{\tilde{\ell}_p(\psi)-\tilde{\ell}_p(\hat{\psi}^{\dagger})\}\left\{\frac{|\tilde{j}_{\lambda\lambda}(\hat{\theta}^{\dagger})|}{|\tilde{j}_{\lambda\lambda}(\hat{\theta}^{\dagger}_{\psi})|}\right\}^{1/2}\frac{\pi(\hat{\theta}^{\dagger}_{\psi})}{\pi(\hat{\theta}^{\dagger})}\{1+O(n^{-3/2})\}, \tag{3.6}$$

where $\hat{\theta}^{\dagger}$ is the MPLE of θ, $\hat{\theta}^{\dagger}_{\psi} = (\psi,\hat{\lambda}^{\dagger}_{\psi})$, with $\hat{\lambda}^{\dagger}_{\psi}$ the constrained MPLE for fixed ψ, $\tilde{j}_p(\psi) = -\partial^2\tilde{\ell}(\theta)/\partial\psi\partial\psi^T|_{\theta=\hat{\theta}^{\dagger}_{\psi}}$ is the pseudo-profile information and $\tilde{j}(\theta) = -\partial^2\tilde{\ell}(\theta)/\partial\theta\partial\theta^T$ is the pseudo-observed information. In the case $p=1$, the corresponding posterior tail

area approximation is readily obtained by integrating (3.6), namely

$$\begin{aligned}
\int_{\psi_0}^{\infty} \tilde{\pi}(\psi|y)\mathrm{d}\psi &= \int_{\psi_0}^{\infty} \frac{|\tilde{j}_p(\hat{\psi})|^{1/2}}{(2\pi)^{1/2}} \exp\{\tilde{\ell}_p(\psi) - \tilde{\ell}_p(\hat{\psi}^\dagger)\} \left\{ \frac{|\tilde{j}_{\lambda\lambda}(\hat{\theta}^\dagger)|}{|\tilde{j}_{\lambda\lambda}(\hat{\theta}^\dagger_\psi)|} \right\}^{1/2} \frac{\pi(\hat{\theta}^\dagger_\psi)}{\pi(\hat{\theta}^\dagger)} \{1 + O(n^{-3/2})\}\, \mathrm{d}\psi \\
&= \int_{\tilde{r}_p(\psi_0)}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{1}{2}\tilde{r}_p(\psi)^2 \right\} \left\{ \frac{\tilde{r}_p(\psi)}{\tilde{q}(\psi)} \right\} \mathrm{d}\tilde{r}_p(\psi) \{1 + O(n^{-3/2})\} \\
&= \Phi\left\{ \tilde{r}_p(\psi_0) + \frac{1}{\tilde{r}_p(\psi_0)} \log \frac{\tilde{q}(\psi_0)}{\tilde{r}_p(\psi_0)} \right\} \{1 + O(n^{-3/2})\} \\
&= \Phi\left\{ \tilde{r}^\star_p(\psi_0) \right\} \{1 + O(n^{-3/2})\}, \qquad (3.7)
\end{aligned}$$

where $\tilde{r}_p(\psi_0) = \mathrm{sign}(\hat{\psi}^\dagger - \psi)[2\{\tilde{\ell}_p(\hat{\psi}^\dagger) - \tilde{\ell}_p(\psi)\}]^{1/2}$ is the pseudo-likelihood root,

$$\tilde{q}(\psi) = \tilde{\ell}_p(\psi) \left\{ \frac{|\tilde{j}_{\lambda\lambda}(\hat{\theta}^\dagger_\psi)|}{|\tilde{j}_{\lambda\lambda}(\hat{\theta}^\dagger)|} \right\}^{1/2} \frac{\pi(\hat{\theta}^\dagger)}{\pi(\hat{\theta}^\dagger_\psi)},$$

and $\tilde{\ell}_p(\psi) = \partial\tilde{\ell}_p(\psi)/\partial\psi$ is the pseudo-profile score. From a practical point of view, the tail area approximation (3.5) ((3.7)) can be used to compute posterior quantiles of ψ, as in Section 2.2.2, but not posterior moments or highest posterior density (HPD) credible intervals. These quantities could in principle be approximated by direct numerical integration of (3.4) ((3.6)). However, for several posterior summaries repeated numerical integrations are needed, and in practice this can be time-consuming. In this section we propose to use the approximate posterior tail area (3.5) (or (3.7)) within the HOTA algorithm (see Sec. 3.1), to produce fast and independent samples from the marginal pseudo-posterior distribution.

3.2.2 Examples

To illustrate the advantages and the accuracy of (3.5) for practical use in Bayesian analyses, we discuss two examples involving pseudo-likelihoods with nuisance parameters. In the following, we consider the HOTA algorithm based on (3.5), *e.g.* based on expansions of the logarithm of the pseudo-posterior. Following the results of Section 3.1, we focus only on the use of (3.5) and the corresponding HOTA algorithm based on its inversion.

The first example focuses on the partial likelihood (see Sec. 2.4.1), usually employed in survival data analysis when the hazard is left unspecified (Cox, 1975). The aim is to approximate the marginal pseudo-posterior of the regression coefficients. In the second example, the pairwise likelihood (see Sec. 2.4.1) is applied to a multivariate normal distribution, with the correlation coefficient being the parameter of interest (Pauli *et al.*, 2011).

For both examples the marginal pseudo-posterior distributions are obtained by inverting the tail area (3.5) via the HOTA algorithm (see Algorithm 4). The proposed approximations are compared also with the random walk Metropolis-Hastings, treated as a gold standard.

Cox regression

To illustrate the higher-order tail approximation to the pseudo-posterior distribution (3.5), we consider a real dataset concerning a clinical study on malignant mesothelioma (MM) Fassina *et al.* (2011). This dataset reports survival times for 109 individuals, with censoring. Moreover the following covariates are provided: type of malignant mesothelioma, i.e. type epithelioid, biphasic or sarcomatoid, gender, epithelial markers (`Cytokeratin`, `E-cadherin`), mesenchymal markers (`N-cadherin`, `vimentin`, `ZEB1`, `ZEB2`, `S100A4`, `MMP2`, `MMP9`, α`-SMA` and `S100A4`). Here we focus on the relation between the covariates and the survival time, so the hazard function has 14 unknown parameters, i.e. $\beta = (\beta_1, \ldots, \beta_{14})$.

The marginal partial posterior distributions for the Cox regression coefficients are approximated by the higher-order asymptotic method implemented with the HOTA algorithm and MCMC, both based on 10^5 final simulations. Here we focus on the

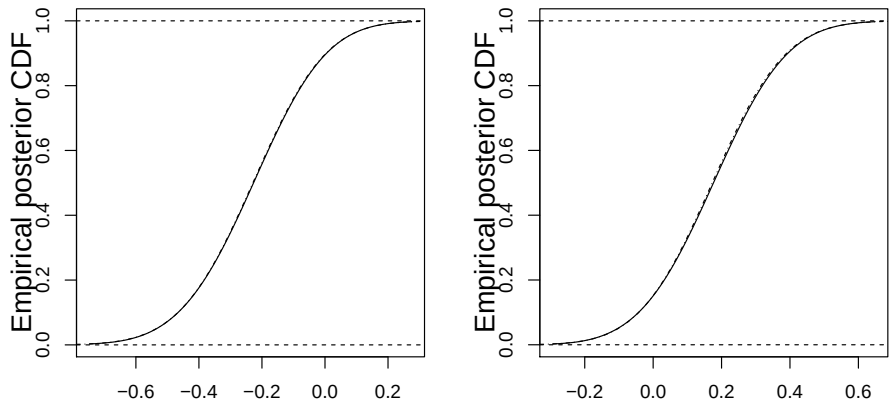


FIGURE 3.6: Cox regression model. Marginal partial posterior distributions for E-cadherin (left) and N-cadherin (right) approximated by HOTA (dot-dashed line) and MCMC (continued).

marginal partial posterior distribution of the effect of E-cadherin (β_4) and N-cadherin (β_5). A graphical comparison of the two methods in terms of the cumulative distribution functions (CDFs) is shown in Figure 3.6 and some numerical comparisons are shown in Table 3.6. Both Figure 3.6 and Table 3.6 highlight the good agreement between MCMC and our method implemented with the HOTA sampling scheme.

Regress. term	Method	Mean	SD	$Q_{2.5}$	Median	$Q_{97.5}$	0.95 HPD
E-cadherin	MCMC	-0.229	0.184	-0.595	-0.227	0.127	(-0.588, 0.131)
	HOTA	-0.230	0.184	-0.594	-0.228	0.125	(-0.601 , 0.117)
N-cadherin	MCMC	0.176	0.170	-0.158	0.177	0.506	(-0.157, 0.507)
	HOTA	0.174	0.169	-0.154	0.173	0.506	(-0.161, 0.499)

TABLE 3.6: Cox regression model. Numerical comparisons of marginal partial posterior distributions.

However, we remark that MCMC produces a dependent sample which is subject to convergence conditions (see Sect. 2.3.1) and which Monte Carlo error may be expensive to reduce. On the other hand, the HOTA algorithm gives an independent sample from the higher-order tail area approximations, and its Monte Carlo error can be controlled essentially without efforts, by simply increasing the simulated values.

Pairwise likelihood

Consider Bayesian inference based on the pairwise likelihood (2.28) obtained from the equi-correlated multivariate normal distribution (see, e.g. Pace *et al.*, 2011). In particular, let Y be a q-variate normal with mean μ, covariance matrix Σ, with $\Sigma_{rr} = \sigma^2$ and $\Sigma_{rs} = \rho\sigma^2$ for $r \neq s$, $r, s = 1, \ldots, q$. In this case the pairwise log-likelihood for $\theta = (\mu, \sigma^2, \rho)$ is given by

$$p\ell(\theta) = -\frac{nq(q-1)}{2}\log\sigma^2 - \frac{nq(q-1)}{4}\log(1-\rho^2) - \frac{q-1+\rho}{2\sigma^2(1-\rho^2)}SS_W - \frac{q(q-1)SS_B + nq(q-1)(\bar{y}-\mu)^2}{2\sigma^2(1+\rho)}, \tag{3.8}$$

where $SS_W = \sum_{i=1}^n \sum_{r=1}^q (y_{ir} - \bar{y}_i)^2$, $SS_B = \sum_{i=1}^n y_{i\bullet}^2$, $\bar{y}_i = \sum_{r=1}^q y_{ir}/q$, and $y_{i\bullet} = q\bar{y}_i$, $i = 1, \ldots, n$.

Following Pauli *et al.* (2011), and given a prior $\pi(\theta)$ we consider the calibrated pairwise posterior

$$\tilde{\pi}_c(\theta|y) \propto \pi(\theta)\exp\{p\ell(\theta)\}^{1/\bar{\omega}}. \tag{3.9}$$

The adjustment $1/\bar{\omega}$ in (3.9) (see Sect. 2.4.1 for its expression) is necessary in order to adjust the curvature of the composite likelihood (Smith & Stephenson, 2009) and allows us to approximately recover the asymptotic properties of the full posterior. To appreciate the relevance of the calibration factor $1/\bar{\omega}$, we consider also the non calibrated pairwise posterior distribution

$$\tilde{\pi}(\theta|y) \propto \pi(\theta)\exp\{p\ell(\theta)\}, \tag{3.10}$$

as proposed by Smith & Stephenson (2009) in the context of spatial extremes.

In this example we focus on the correlation parameter, *e.g.* $\psi = \rho$. Since the full likelihood is analytically available, then the pairwise posterior is compared also with the full posterior (see Figure 3.7 and Table 3.7). A vague normal prior is assumed for $\left(\mu, \log\sigma^2, \Phi^{-1}\left\{\frac{(q-1)\rho+1}{q}\right\}\right)$, with independent components. A sample of size $n = 10$ is considered from the standard equi-correlated q-variate normal distribution, with $\rho = 0.5$ and $q = 20$. The marginal posterior distributions for ρ are approximated by MCMC and the HOTA algorithm, both based on 10^5 random draws.

Figure 3.7 compares three marginal posteriors of ρ. In particular it shows the full marginal posterior, the marginal pairwise-posterior obtained from (3.10) and the adjusted marginal pairwise-posterior based on (3.9), all approximated with HOTA and MCMC. Moreover, the boxplots give a comparison of the three posteriors computed with HOTA. The two approximation methods give very similar results, and this is essentially confirmed also by the summary statistics reported in Table 3.7.

Lastly, the boxplots highlight that Bayesian inference based on (3.10) is falsely precise (see also Pauli *et al.*, 2011)

Posterior	Method	Mean	SD	$Q_{0.025}$	Median	$Q_{0.975}$	0.95 HPD
Adj. pair	HOTA	0.568	0.133	0.305	0.571	0.813	(0.311, 0.818)
Adj. pair	MCMC	0.570	0.135	0.299	0.574	0.814	(0.316, 0.827)
Full	HOTA	0.562	0.119	0.342	0.558	0.801	(0.34, 0.798)
Full	MCMC	0.564	0.119	0.342	0.562	0.803	(0.342, 0.803)
Pair	HOTA	0.518	0.017	0.485	0.518	0.551	(0.486, 0.551)
pair	MCMC	0.519	0.017	0.485	0.519	0.551	(0.485, 0.551)

TABLE 3.7: Equi-correlated normal model. Summaries of the full, pairwise and adjusted pairwise posterior distribution approximated by HOTA and MCMC.

3.2.3 Remarks

By paralleling results for genuine posterior distributions, we discussed higher-order approximations for pseudo-posterior distributions, *i.e.* posterior distribution based on pseudo-likelihood functions. This theory provides asymptotic formulae for tail area and posterior quantiles, which are available at little additional cost over simple first-order approximations.

Moreover, these approximations can be easily implemented through the HOTA sampling scheme (see Sect. 3.1) to approximate marginal pseudo-posterior densities and posterior summaries very quickly. Finally, the proposed method combined with HOTA inherits all the advantages of the latter, hence it provides a convenient framework for quick prior sensitivity analyses (Reid & Sun, 2010).

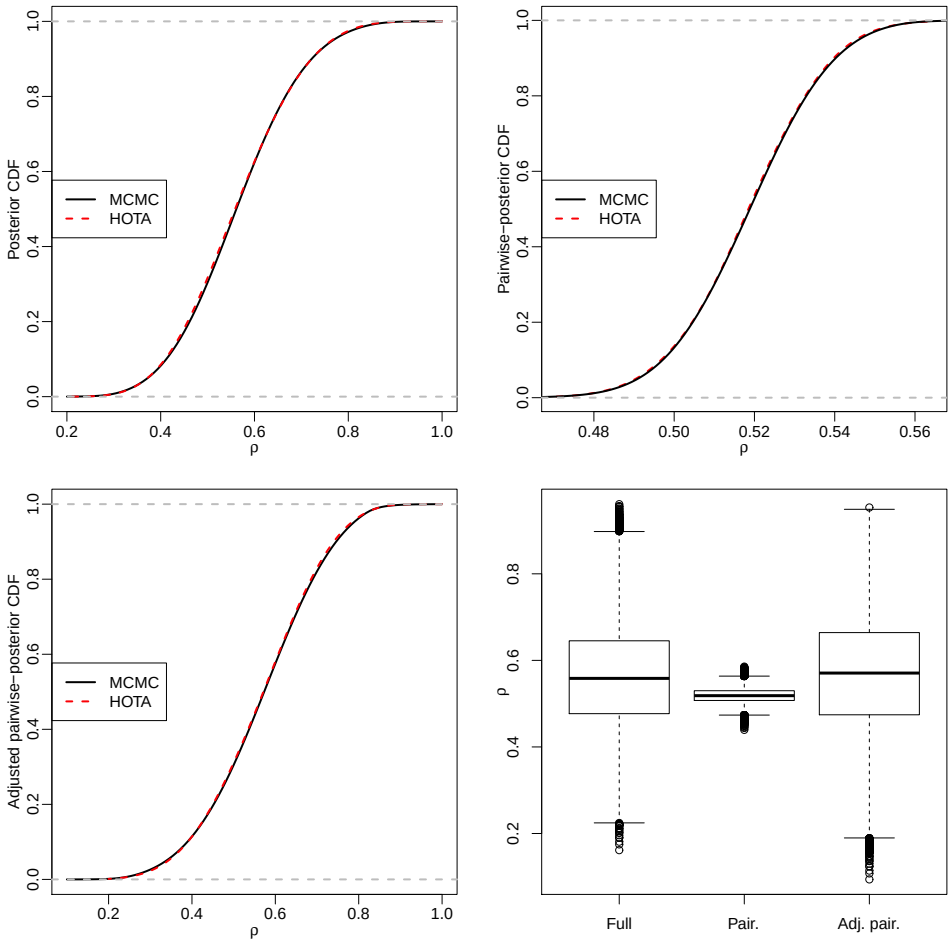


FIGURE 3.7: Equi-correlated normal model. Full (top-left), pairwise (top-right) and adjusted pairwise (bottom-left) marginal posteriors of ρ, approximated by HOTA and MCMC. The boxplots (bottom-right) compare the three marginal posteriors computed with HOTA.

3.3 Approximate credible sets via modified log-likelihood ratios

Approximate credible intervals for a scalar a parameter of interest based on modifications of the likelihood root, such as (2.13) and (2.14), have been widely discussed in the Bayesian literature; see, among others DiCiccio *et al.* (1990); Sweeting (1995, 1996, 1999); Ventura *et al.* (2013).

Consider the posterior distribution (2.1) with θ scalar. Then the modified likelihood root function is (Sweeting, 1996; Ventura *et al.*, 2013)

$$r^{\star}(\theta) \;=\; r(\theta) + \frac{1}{r(\theta)} \log \frac{q(\theta)}{r(\theta)}, \tag{3.11}$$

where $r(\theta) = \text{sign}(\hat{\theta} - \theta)[2\{\ell(\hat{\theta}) - \ell(\theta)\}]^{1/2}$ and $q(\theta) = \ell'(\theta)\pi(\hat{\theta})/\{j(\hat{\theta})^{1/2}\pi(\theta)\}$. It can be shown that (3.11) is asymptotically standard normal with error of order $O(n^{-3/2})$.

The modified likelihood root (3.11) can be obtained by following the three step procedure discussed in Skovgaard (2001); see also Davison (2003, Ch. 11) and references therein.

Step 1: Consider the Laplace expansion of $\pi(\theta|y)$, given by

$$\pi(\theta|y) \;=\; \frac{1}{\sqrt{2\pi}}|j(\hat{\theta})|^{1/2}\frac{\pi(\theta)}{\pi(\hat{\theta})}\exp\left\{-\frac{1}{2}r(\theta)^2\right\}\{1+O(n^{-1})\}\,, \qquad (3.12)$$

Step 2: Change the variable from θ to $r = r(\theta)$. A motivation for considering such a transformation is that, in terms of r^2, the quantity $\exp(-r^2/2)$ in (3.12) is the kernel of the standard normal density. The Jacobian is $dr(\theta)/d\theta = -\ell'(\theta)/r(\theta)$, and thus

$$\pi(r|y) \;=\; \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}r^2 + \log b(r)\right\}\{1+O(n^{-1})\}\,,$$

where the positive quantity $b(r) = |j(\hat{\theta})|^{1/2}\frac{\pi(\theta)}{\pi(\hat{\theta})}\frac{r(\theta)}{\ell'(\theta)}$ is regarded as a function of r.

Step 3: Change of variable from r to $r^* = r^*(\theta) = r - r^{-1}\log b(r)$, so that

$$-(r^*)^2 = -r^2 + 2\log b(r) - \left(r^{-1}\log b(r)\right)^2. \qquad (3.13)$$

The Jacobian of the transformation and the third term in (3.13) contribute only to the error, and it can be shown that (see Sweeting, 1995, 1996, Severini, 2000, Ch. 2)

$$\pi(r^*|y) \;=\; \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(r^*)^2\right\}\{1+O(n^{-3/2})\}. \qquad (3.14)$$

Note that from (3.14) the following tail area approximation can be derived

$$\begin{aligned}\int_{\theta_0}^{\infty}\pi(\theta|y)\,d\theta \;&=\; \frac{1}{\sqrt{2\pi}}\int_{r_0^*}^{\infty}\exp\left\{-\frac{1}{2}(r^*)^2\right\}dr^*\{1+O(n^{-3/2})\}\\ &=\; \Phi(r_0^*)\{1+O(n^{-3/2})\}, \qquad (3.15)\end{aligned}$$

where $r_0^* = r^*(\theta_0)$. Formula (3.15) gives an explicit expression for the posterior quantiles. Moreover, as seen in Section 3.1, (3.15) gives rise to the HOTA simulation scheme for approximate marginal posterior simulation.

From (3.15) an approximate credible interval for θ can be computed as $CI_{1-\alpha} = \{\theta : w^*(\theta) \le \chi^2_{1,1-\alpha}\}$, where $w^*(\theta) = r^*(\theta)^2$ and $\chi^2_{1,1-\alpha}$ is the $(1-\alpha)$-quantile of the χ^2_1

distribution. Equivalently, $CI_{1-\alpha}$ can be computed as

$$CI_{1-\alpha} = \left\{\theta : |r^*(\theta)| \leq z_{1-\alpha/2}\right\}. \tag{3.16}$$

Note that (3.16) defines a third-order equi-tailed credible interval for θ with accurate sampling coverage (see Sweeting, 1999).

It is of interest to extend the theory of asymptotic expansions for a vector parameter of interest. As is the case with the approximations for a scalar parameter, the proposed results are based on the asymptotic theory of modified log-likelihood ratios (Skovgaard, 2001), they require only routine maximization output for their implementation, and they are constructed for arbitrary prior distributions. Moreover, the proposed results are analytical and do not require simulation from the posterior distribution. From a practical point of view, the asymptotic expansions can be used to compute approximate Bayesian credible sets with accurate posterior probability content and sampling coverage. These credible sets can be seen as a multivariate generalization of the equi-tailed credible interval (3.16).

3.3.1 Modified log-likelihood ratios

Suppose that $\theta \in \Theta \subseteq \mathbb{R}^d$, with $d > 1$. It is possible to extend the three-step procedure suggested above to posterior distributions with d parameters as follows (see Skovgaard, 2001, for a frequentist extension). Consider the following three steps:

Step 1: compute the Laplace approximation of $\pi(\theta|y)$, given by

$$\pi(\theta|y) \;=\; (2\pi)^{-d/2}|j(\hat{\theta})|^{1/2}\exp\left\{-\frac{1}{2}w(\theta)\right\}\frac{\pi(\theta)}{\pi(\hat{\theta})}\{1+O(n^{-1})\},$$

where $w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\}$;

Step 2: change the variable of integration from θ to $r_m = r_m(\theta)$, such that for the log-likelihood ratio we have $w(\theta) = 2\{\ell(\hat{\theta}) - \ell(\theta)\} = r_m(\theta)^T r_m(\theta)$;

Step 3: change the variable of integration from r_m to a perturbed version of the form $r_m^* = r_m^*(\theta) = r_m - \delta(r_m)$, with $\delta = \delta(r_m)$ chosen to satisfy $r_m^T\delta(r_m) = \log g(r_m)$ for a suitably defined term $g(r_m)$, so that $(r_m-\delta)^T(r_m-\delta) = r_m^T r_m - 2\log g(r_m) + O(n^{-2})$ is asymptotically χ^2_d.

In order to compute Step 2, we need a statistic $r_m = r_m(\theta)$ for which $r_m^T r_m = w(\theta)$. To this end let us consider the signed root log-likelihood ratio transformation defined in Sweeting (1995, 1996); see also Kharroubi & Sweeting (2010). In particular, let

$\theta = (\theta_1, \ldots, \theta_d) = (\theta_{1:i}, \theta^{i+1:d})$, where $\theta_{1:i} = (\theta_1, \ldots, \theta_i)$ is the vector of the first i components of θ and $\theta^{i+1:d} = (\theta_{i+1}, \ldots, \theta_d)$. To state notation, let $\hat{\theta}^{i+1:d}_{\theta_{1:i}}$ be the partial MLE of $\theta^{i+1:d}$ given $\theta_{1:i}$, and let $\hat{\theta}_{j,\theta_{1:i}}$ be the jth component of $(\theta_{1:i}, \hat{\theta}^{i+1:d}_{\theta_{1:i}})$, for $j > i$. The signed root log-likelihood ratio transformation is thus defined as

$$r_m(\theta) = (r_{m1}, \ldots, r_{md}) \ , \tag{3.17}$$

with

$$r_{mi} = \text{sign}(\theta_i - \hat{\theta}_{i,\theta_{1:i-1}}) \left[2 \left\{ \ell \left(\theta_{1:i-1}, \hat{\theta}^{i:d}_{\theta_{1:i-1}} \right) - \ell \left(\theta_{1:i}, \hat{\theta}^{i+1:d}_{\theta_{1:i}} \right) \right\} \right]^{1/2} . \tag{3.18}$$

Note that (3.18) is a function of $\theta_{1:i}$. Moreover, $r_m(\theta)$ is a one-to-one data-dependent transformation of θ, such that $\exp\left\{-\frac{1}{2} r_m^T r_m\right\} = L(\theta)/L(\hat{\theta})$. Finally, $r_m(\theta)$ is asymptotically multivariate standard normal to $O(n^{-1/2})$ (Sweeting, 1995).

In the second step, when changing the variable of integration from θ to the statistic r_m, given in (3.17), the Jacobian matrix $dr_m/d\theta$ is lower triangular (Kharroubi & Sweeting, 2010)

$$\left| \frac{dr_m}{d\theta} \right| = \prod_{i=1}^{d} \left| \frac{\ell_i \left(\theta_{1:i}, \hat{\theta}^{i+1:d}_{\theta_{1:i}} \right)}{r_{mi}} \right| ,$$

where $\ell_i(\theta)$ is the ith component of the score vector $\partial \ell(\theta)/\partial\theta$, for $i = 1, \ldots, d$.

In the last step, following Skovgaard (2001), we perturb r_m to $r_m^* = r_m^*(\theta) = r_m - \delta(r_m)$, with $\delta(r_m)$ chosen to satisfy $r_m^T \delta(r_m) = \log g(r_m)$, so that

$$-\{r_m - \delta(r_m)\}^T \{r_m - \delta(r_m)\} = -r_m^T r_m + 2 \log g(r_m) + O(n^{-2}) \ . \tag{3.19}$$

In order to compute (3.19), we only need the existence of $\delta(r_m)$ to calculate

$$w_m^* = w_m^*(\theta) = r_m(\theta)^T r_m(\theta) - 2 \log g(r_m(\theta)) \ , \tag{3.20}$$

with

$$g(r_m(\theta)) = |j(\hat{\theta})|^{1/2} \frac{\pi(\theta)}{\pi(\hat{\theta})} \left\{ \prod_{i=1}^{d} \left| \frac{\ell_i \left(\theta_{1:i}, \hat{\theta}^{(i+1)}_{\theta_{1:i}} \right)}{r_{mi}} \right| \right\}^{-1} . \tag{3.21}$$

The asymptotic distribution of w_m^* is χ^2_d with relative error $O(n^{-1})$ in a large deviation region. An asymptotically equivalent version suggested by Skovgaard (2001) is

$$w_m^{**} = w_m^{**}(\theta) = r_m^T r_m \left(1 - \frac{\log g(r_m)}{r_m^T r_m} \right)^2 . \tag{3.22}$$

From (3.22), or (3.20), an approximate credible set for θ can be computed as

$$CR = \left\{\theta : w^{**}(\theta) \leq \chi^2_{d,1-\alpha}\right\}, \tag{3.23}$$

where $\chi^2_{d,1-\alpha}$ is the $1-\alpha$th quantile of the χ^2_d distribution with d degrees of freedom. This credible region has $100(1-\alpha)\%$ coverage in repeated sampling with relative error $O(n^{-1})$ in a large deviation region, and thus improves upon the credible region based on the normal approximation (see also Sect. 2.2.1.)

$$CR_N = \left\{\theta : (\theta - \tilde{\theta})^T \tilde{V} (\theta - \tilde{\theta}) \leq \chi^2_{d,1-\alpha}\right\}, \tag{3.24}$$

and the likelihood-type credible region

$$CR_L = \left\{\theta : -2 \log \frac{\pi(\theta|y)}{\pi(\tilde{\theta}|y)} \leq \chi^2_{d,1-\alpha}\right\} . \tag{3.25}$$

Note that, in general, the credible set (3.24) may be questionable since it is based on the normal approximation, which forces credible sets to have an elliptical shape.

3.3.2 Examples

In this section the accuracy of (3.23) is illustrated empirically by means of three examples. From a Bayesian perspective, we check the posterior probability content of the credible regions by simulating values from the posterior distributions via MCMC methods, whereas from the frequentist perspective we check the empirical coverage of the suggested credible sets under repeated sampling from the model with a fixed parameter value. We do not pursue the comparison with (3.20) as it gave very similar results to (3.23).

Normal distribution

Consider a random sample $y = (y_1, \ldots, y_n)$ from a $N(\mu, \sigma^2)$ distribution, with $\theta = (\mu, \sigma^2)$ unknown. We assume two different prior distributions of θ, *i.e.*, the improper prior $\pi_1(\theta) \propto 1/\sigma^2$ and an informative normal-inverse gamma prior $\pi_2(\theta)$. In this situation, all the quantities involved in the computation of w^* and w^{**} are easy to compute.

For a sample of size $n = 10$, Figure 3.8 shows several credible regions for $\theta = (\mu, \sigma^2)$ with the improper prior, i.e. CR_N, CR_L, the CR region based on w^{**}_m and the exact 95% HPD credible region computed by simulation. For each credibility region we compute the posterior probability content as the proportion of simulated values laying inside the

	$\pi_1(\theta)$			$\pi_2(\theta)$		
$1-\alpha$	0.90	0.95	0.99	0.90	0.95	0.99
	$n=10$			$n=10$		
CR_N	0.7280	0.7830	0.8685	0.5905	0.6470	0.7402
CR_L	0.8540	0.9130	0.9770	0.7871	0.8688	0.9578
CR	0.9075	0.9510	0.9925	0.9020	0.9517	0.9904
	$n=15$			$n=15$		
CR_N	0.7615	0.8280	0.900	0.6698	0.7302	0.8189
CR_L	0.8485	0.9225	0.984	0.8276	0.8992	0.9738
CR	0.8935	0.9500	0.990	0.9050	0.9544	0.9916
	$n=30$			$n=30$		
CR_N	0.8275	0.889	0.9495	0.7688	0.8250	0.9031
CR_L	0.8775	0.936	0.9840	0.8606	0.9242	0.9824
CR	0.8980	0.948	0.9875	0.9023	0.9533	0.9888
	$n=50$			$n=50$		
CR_N	0.8630	0.9240	0.9730	0.8160	0.8761	0.9436
CR_L	0.8965	0.9435	0.9890	0.8791	0.9346	0.9836
CR	0.9045	0.9525	0.9890	0.9011	0.9514	0.9897

TABLE 3.8: Normal distribution. Empirical coverage probabilities of credible regions.

defined region. In this example, the posterior probability content is 0.674 for CR_N, 0.881 for CR_L and 0.949 CR. Only CR has the correct posterior probability.

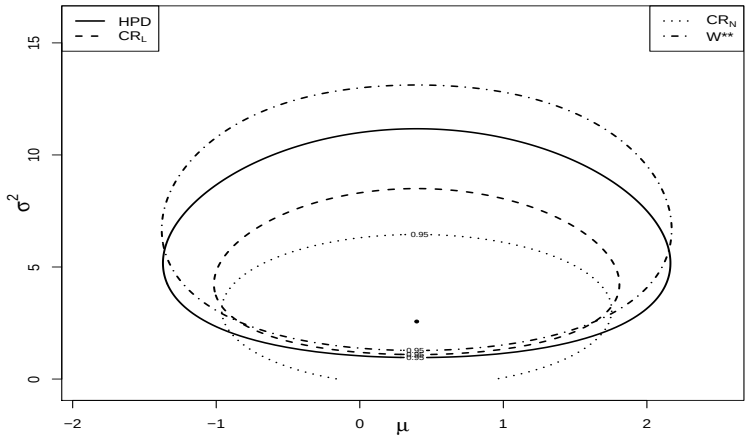


FIGURE 3.8: Normal distribution. Credible regions for (μ, σ^2) with the improper prior.

To judge the sampling properties of the credible region (3.23), we check the empirical coverage probability in a simulation study based on 10^4 Monte Carlo trials. Table 3.8 gives the empirical coverages for $(1-\alpha)$ posterior credible regions (3.23) in comparison to the credible regions CR_N and CR_L. From Table 3.8 we note that, for every n, CR clearly improves on (3.24) and (3.25). Larger sample sizes (not reported here) show, as one would expect, rather little differences between the results of all the procedures.

Gamma distribution

Consider a random sample $y = (y_1, \ldots, y_n)$ from a gamma distribution, with both the shape κ and scale σ parameters unknown. Let us consider $\theta = (\log \sigma, \log \kappa)$ and for θ we assume two prior distributions, that are $\pi_1(\theta) \propto 1$ and $\pi_2(\theta) = N(\mu, \nu)^2$, where (μ, ν) is a fixed hyperparameter.

As in the previous example, to judge the coverage quality of CR, a simulation study based on 2000 Monte Carlo trials has been performed. Table 3.9 gives the empirical coverage probabilities for (3.23) in comparison to the first-order credible regions CR_N and CR_L.

	π_1			$\pi_2(\mu = 0, \nu = 10)$			$\pi_2(\mu = 3, \nu = 10)$		
$1-\alpha$	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
	$n = 5$			$n = 5$			$n = 5$		
CR_N	0.8188	0.7991	0.8801	0.7642	0.8288	0.9040	0.6630	0.7324	0.8374
CR_L	0.8405	0.9084	0.9755	0.8624	0.9265	0.9837	0.7787	0.8659	0.9594
CR	0.9018	0.9500	0.9895	0.9166	0.9612	0.9933	0.8753	0.9338	0.9864
	$n = 10$			$n = 10$			$n = 10$		
CR_N	0.8188	0.8779	0.9445	0.8281	0.8868	0.9495	0.7764	0.8381	0.9215
CR_L	0.8748	0.9336	0.9832	0.8826	0.9385	0.9854	0.8402	0.9115	0.9764
CR	0.9028	0.9519	0.9893	0.9084	0.9564	0.9908	0.8866	0.9424	0.9872

TABLE 3.9: Gamma model. Empirical coverage probabilities of credible regions.

From Table 3.9 we note that, for every n, CR improves on (3.24) and (3.25). Observe also that for parameter values in regions of low prior density there may be, as expected, some degradation in the coverage accuracy.

Weibull regression

Let us consider a random sample $(t_1, \ldots, t_n)$ from the Weibull distribution with shape parameter κ and scale parameter $\lambda_i = x_i^T \beta$, where x_i is a $p \times 1$ vector of covariates, $i = 1, \ldots, n$, and the unknown parameters are the $p \times 1$ vector β and κ. Note that $y_i = \log t_i$ follows a regression and scale model of the form $y_i = x_i^T \beta + \sigma \varepsilon_i$, with $\sigma = 1/\kappa$ and ε_i log-Weibull or extreme-value random variables, $i = 1, \ldots, n$.

For the parameter $\theta = (\log \sigma, \beta)$ we assume two prior distributions, *i.e.* the noninformative prior $\pi_1(\theta) \propto 1$ and the proper prior $\pi_2(\theta) = \prod_{i=1}^{p+1} N(\mu_i, 20)$, where $\mu = (\mu_1, \ldots, \mu_{p+1})$ is a fixed hyperparameter.

A simulation study based on 5000 Monte Carlo trials has been performed with $p = 4$ and $p = 9$ in order to judge the coverage quality of CR in comparison to the first-order credible regions CR_N and CR_L.

	$p=4$						$p=9$					
	π_1			π_2			π_1			π_2		
$1-\alpha$	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
	$n=10$			$n=10$			$n=15$			$n=15$		
CR_N	0.4292	0.5054	0.6176	0.4530	0.5282	0.6424	0.0128	0.1610	0.2332	0.1380	0.1780	0.2540
CR_L	0.7322	0.8270	0.9374	0.7490	0.8432	0.9448	0.5008	0.6094	0.7992	0.5226	0.6314	0.8132
CR	0.9348	0.9700	0.9944	0.9424	0.9754	0.9962	0.9560	0.9776	0.9966	0.9634	0.9828	0.9970
	$n=20$			$n=20$			$n=20$			$n=20$		
CR_N	0.6736	0.7444	0.8496	0.6814	0.7528	0.8550	0.2656	0.3300	0.4584	0.2758	0.3332	0.4548
CR_L	0.8382	0.9114	0.9774	0.8452	0.9154	0.9792	0.6582	0.7628	0.9034	0.6592	0.7650	0.9100
CR	0.9190	0.9592	0.9920	0.9234	0.9636	0.9932	0.9472	0.9750	0.9954	0.9528	0.9796	0.9964
	$n=30$			$n=30$			$n=30$			$n=30$		
CR_N	0.7526	0.8242	0.9114	0.7564	0.8282	0.9130	0.4332	0.5134	0.6536	0.4580	0.5378	0.6694
CR_L	0.8616	0.9238	0.9794	0.8650	0.9262	0.9806	0.7562	0.8480	0.9494	0.7692	0.8516	0.9540
CR	0.9074	0.9534	0.9912	0.9104	0.9558	0.9914	0.9356	0.9686	0.9948	0.9388	0.9708	0.9954
	$n=50$			$n=50$			$n=50$			$n=50$		
CR_N	0.8058	0.8756	0.9454	0.8088	0.8768	0.9464	0.6286	0.7090	0.8290	0.6178	0.7006	0.8254
CR_L	0.8762	0.9336	0.9864	0.8784	0.9360	0.9870	0.8294	0.9044	0.9716	0.8254	0.9032	0.9756
CR	0.9052	0.9530	0.9926	0.9072	0.9534	0.9926	0.9220	0.9616	0.9928	0.9220	0.9608	0.9918

TABLE 3.10: Weibull regression model. Empirical coverage probabilities of credible regions; the hyperparameter μ is fixed equal to the true parameter values $(\log 2, -1, 1, -1, 1)$ for $p = 4$ and to $(\log 2, -1, 1, -1, 1, -1, 1, -1, 1, -1)$ for $p = 9$.

From Table 3.10 we note that, for every n and p, CR is always preferable to (3.24) and (3.25).

3.3.3 Remarks

We have shown that modified log-likelihood ratios provide important quantities useful to obtain approximate Bayesian credible regions for a vector parameter, with very little computational effort and in a fraction of the time required for a full simulation approach. Although the approximations described in this section are derived from asymptotic considerations, they perform extremely well in small sample situations.

A key feature of the approximations discussed and developed in this section is that they require only the calculation of the first and second derivative of the log-likelihood function, as well as likelihood maximizations.

Finally, note that the signed root log-likelihood ratio transformation (3.17) in general depends on the chosen parameter order. However, in the examples considered in the previous section, the results of the simulation studies do not change (results not reported here) when inverting the parameter order.

The credible regions suggested here may be used also outside the Bayesian setting, as accurate confidence sets. Indeed, as shown by means of empirical coverages the suggested method does produce such regions, which improve significantly over usual likelihood-based or Wald-type regions.

3.4 An improved Laplace approximation for marginal likelihoods

Bayes factors (see Sect. 2.1), which are typically used for Bayesian model selection, are based on posterior model probabilities or marginal likelihoods. In practice the computation of such marginal likelihoods may be challenging and typically it requires more effort than the usual posterior sampling.

The Laplace approximation (2.7) may be used for this purpose, but its accuracy may be questionable, especially in small samples. DiCiccio *et al.* (1997) show that the accuracy of the Laplace approximation may be improved via a Bayesian Bartlett correction, which in general requires posterior simulation. Another popular method for computing marginal likelihoods is suggested by Chib (1995), where the posterior simulation is done with the Gibbs sampling, and extended by Chib & Jeliazkov (2001) using posterior simulation via Metropolis-Hastings (see also Section 2.4.2). However, implementing both methods requires a careful partitioning of the parameter. In addition it requires a considerable amount of bookkeeping.

In this section we show how the identity (2.22) used by Chib (1995) and Chib & Jeliazkov (2001) can be exploited alongside the Laplace approximation for marginal distributions (Tierney & Kadane, 1986), to obtain an improved Laplace-type approximation for $p(y)$, called HOA-Laplace. We show, both theoretically and empirically by means of numerical examples, that this approximation has asymptotic error of order $O(n^{-3/2})$, which is superior to the usual Laplace approximation.

Although the proposed method is theoretically less accurate than the Bartlett-corrected version (2.17), it has the remarkable advantage of not requiring posterior simulation. Another generalization of the Laplace approximation is proposed in Nott *et al.* (2009), which attempts to find changes of variable for which the integrand becomes approximately a product of one-dimensional functions. However, as it will be shown in the examples the method of Nott *et al.* (2009) is less accurate then HOA-Laplace, both theoretically and empirically.

From a practical point of view, HOA-Laplace requires only nested posterior maximizations, evaluation of the posterior Hessian and univariate numerical integrations, all of which can be easily obtained with software such as R (R Core Team, 2013).

3.4.1 Background and theory

Assume that $L(\theta)$ satisfies the usual regularity conditions required for the Laplace approximation to be valid (see Tierney & Kadane, 1986; Kass *et al.*, 1990). The posterior distribution can be written as

$$\pi(\theta|y) \;=\; \pi(\theta_1|y)\pi(\theta_2|\theta_1 y)\cdots\pi(\theta_d|\theta_1,\ldots,\theta_{d-1},y). \tag{3.26}$$

If all the factors on the right-hand side of (3.26) are known, then, given a point θ^*, $p(y)$ can be readily computed from the identity (2.22) as

$$\log\hat{p}(y) = h(\theta^*) - \log\pi(\theta_1^*|y) - \log\pi(\theta_2^*|\theta_1^*,y) - \cdots - \log\pi(\theta_d^*|\theta_1^*,\ldots,\theta_{d-1}^*,y). \tag{3.27}$$

However, this is not possible in general since some factors in (3.26) may not have a closed form expression, and their computation my require awkward multidimensional integrals.

The Laplace approximation for marginal posterior distributions (Tierney & Kadane, 1986) is a useful tool for separating the components of $\pi(\theta|y)$. The idea behind the proposed HOA-Laplace method is to approximate each component in (3.26) via the Laplace method, and then renormalize it numerically in order to gain accuracy.

In particular, let $\theta = (\theta_{1:i}, \theta^{i+1:d})$, $i = 1,\ldots,d-1$. Formula (3.26) can be rewritten as

$$\frac{\int\exp\{h(\theta)\}d\theta_{2:d}}{\int\exp\{h(\theta)\}d\theta}\frac{\int\exp\{h(\theta)\}d\theta_{3:d}}{\int\exp\{h(\theta)\}d\theta_{2:d}}\cdots\frac{\exp\{h(\theta)\}}{\int\exp\{h(\theta)\}d\theta_d}. \tag{3.28}$$

Let us denote by $\tilde{\theta}^{i+1:d}_{\theta_{1:i}}$ the mode of the unormalized log-posterior $h(\theta)$ with $\theta_{1:i}$ fixed, and let $\tilde{\theta}^{j+1:d}_{\theta^*_{1:j-1}\theta_j}$ be the mode of $h(\theta)$ with $\theta^*_{1:j-1} = (\theta_1^*,\ldots,\theta_{j-1}^*)$ and θ_j fixed.

The first ratio in (3.28), *e.g.* the marginal posterior of θ_1, can be approximated via the Laplace method as (2.10), *i.e.*

$$\hat{\pi}_L(\theta_1|y) \;\propto\; \exp\left\{h(\theta_1,\tilde{\theta}^{2:d}_{\theta_1})\right\}\left|V_{\theta^{2:d}\theta^{2:d}}(\theta_1,\tilde{\theta}^{2:d}_{\theta_1})\right|^{-1/2}, \tag{3.29}$$

where $V_{\theta^{2:d}\theta^{2:d}}(\theta)$ is the block $(\theta^{2:d}\theta^{2:d})$ of the posterior information matrix $V(\theta)$. Following Tierney & Kadane (1986) and Tierney *et al.* (1989), the renormalized Laplace approximation is third-order accurate, *e.g.*

$$\begin{aligned}\pi(\theta_1|y) &= \frac{\hat{\pi}_L(\theta_1|y)}{\int\hat{\pi}_L(\theta_1|y)\,d\theta_1}\{1+O(n^{-3/2})\}\\ &= \hat{\pi}^*_L(\theta_1|y)\{1+O(n^{-3/2})\},\end{aligned} \tag{3.30}$$

where the integration in the denominator can be performed numerically, via any quadrature rule. Evaluating $\pi_L^*(\theta_1|y)$ at θ_1^* gives a third-order approximation to the posterior marginal density $\pi(\theta_1^*|y)$.

The second ratio in (3.28), namely $\pi(\theta_2|\theta_1, y)$, can be approximated by following essentially the same line of reasoning, with the only difference in that θ_1 is fixed to θ_1^*. In other word we seek and approximation of $\pi(\theta_2|\theta_1^*, y)$, which is a univariate density. Hence, the Laplace approximation to this conditional posterior density is

$$\hat{\pi}_L(\theta_2|\theta_1^*, y) \propto \exp\left\{h(\theta_1^*, \theta_2, \tilde{\theta}^{3:d}_{\theta_1^*\theta_2}) - h(\theta_1^*, \tilde{\theta}^{2:d}_{\theta_1^*})\right\} \left\{ \frac{|V_{\theta^{2:d}\theta^{2:d}}(\theta_1^*, \tilde{\theta}^{2:d}_{\theta_1^*})|}{|V_{\theta^{3:d}\theta^{3:d}}(\theta_1^*, \theta_2, \tilde{\theta}^{3:d}_{\theta_1^*\theta_2})|} \right\}^{1/2}, \quad (3.31)$$

and its renormalized version is

$$\begin{aligned} \pi(\theta_2|\theta_1^*, y) &\stackrel{a}{=} \frac{\hat{\pi}_L(\theta_2|\theta_1^*, y)}{\int \hat{\pi}_L(\theta_2|\theta_1^*, y)\, d\theta_2} \\ &\stackrel{a}{=} \hat{\pi}_L^*(\theta_2|\theta_1^*, y), \end{aligned} \quad (3.32)$$

where the symbol "$\stackrel{a}{=}$" means equality for $n \to \infty$. An approximation to the required conditional posterior density is then given by $\hat{\pi}_L^*(\theta_2^*|\theta_1^*, y)$.

More generally, for $j = 2, \ldots, d-1$, the unnormalized Laplace approximation to the conditional posterior densities is

$$\begin{aligned} \hat{\pi}_L(\theta_j|\theta^*_{1:j-1}, y), \quad \propto \quad & \exp\left\{h(\theta^*_{1:j-1}, \theta_j, \tilde{\theta}^{j+1:d}_{\theta^*_{1:j-1}\theta_j}) - h(\theta^*_{1:j-1}, \tilde{\theta}^{j:d}_{\theta^*_{1:j-1}})\right\} \\ & \left\{ \frac{|V_{\theta^{j:d}\theta^{j:d}}(\theta^*_{1:j-1}, \tilde{\theta}^{j:d}_{\theta^*_{1:j-1}})|}{|V_{\theta^{j+1:d}\theta^{j+1:d}}(\theta^*_{1:j-1}, \theta_j, \tilde{\theta}^{j+1:d}_{\theta^*_{1:j-1}\theta_j})|} \right\}^{1/2}, \end{aligned} \quad (3.33)$$

and the corresponding normalized version is denoted by

$$\hat{\pi}_L^*(\theta_j|\theta^*_{1:j-1}, y) = \frac{\hat{\pi}_L(\theta_j|\theta^*_{1:j-1}, y)}{\int \hat{\pi}_L(\theta_j|\theta^*_{1:j-1}, y)\, d\theta_j}, \quad (3.34)$$

The last conditional posterior density is

$$\pi(\theta_d|\theta^*_{1:d-1}, y) \;=\; \frac{\exp\left\{h(\theta^*_{1:d-1}, \theta_d)\right\}}{\int \exp\left\{h(\theta^*_{1:d-1}, \theta_d\right\}\, d\theta_d}. \quad (3.35)$$

and thus it can be computed exactly.

As all the integrals required for the renormalization of the suggested approximation are univariate, they can easily be approximated via any quadrature rule, with extreme accuracy (by, *e.g.* , the `integrate` function in R).

The product of the approximate marginal posterior (3.30) with the product of the approximate conditionals $\prod_{j=2}^{d-1} \hat{\pi}_L^*(\theta_j|\theta_{1:j-1}^*, y)$ and (3.35), all evaluated at θ^*, gives an approximation to the joint posterior density at θ^*. Let us denote this approximate posterior density by

$$\pi_L^*(\theta^*|y) = \hat{\pi}_L^*(\theta_1^*|y)\hat{\pi}_L^*(\theta_2^*|\theta_1^*, y)\cdots\pi(\theta_d^*|\theta_1^*, \ldots, \theta_{d-1}^*, y),. \tag{3.36}$$

The idea of HOA-Laplace approximation is to replace the posterior densities required in (3.27) by their approximate version given by (3.36), which leads to an approximate marginal likelihood accurate to $O(n^{-3/2})$.

To see why the method has third-order accuracy, note that $\pi(\theta_1^*|y) = \hat{\pi}_L^*(\theta_1^*|y)\{1 + O(n^{-3/2})\}$; see, for instance, Tierney & Kadane (1986) and Tierney *et al.* (1989). Moreover, since the conditional posterior density of θ_j given the previous components – fixed at $\theta_{1:j-1}^*$ – is just a marginal density, it is easily seen that $\pi(\theta_j|\theta_{1:j-1}^*, y) = \hat{\pi}_L^*(\theta_j|\theta_{1:j-1}^*)\{1 + O(n^{-3/2})\}$, with $j = 2, \ldots, d-1$. As the last conditional density in (3.36) is computed exactly, we have that

$$\begin{aligned}\pi(\theta^*|y) &= \hat{\pi}_L^*(\theta_1^*|y)\hat{\pi}_L^*(\theta_2^*|\theta_1^*, y)\cdots\pi(\theta_d^*|\theta_1^*, \ldots, \theta_{d-1}^*, y)\{1 + O(n^{-3/2})\}^{d-1} \\ &= \hat{\pi}_L^*(\theta^*|y)\{1 + O(n^{-3/2})\},\end{aligned}$$

which implies that $p(y) = p_L^*(y)\{1 + O(n^{-3/2})\}$.

To show numerically that HOA-Laplace is third-order accurate we study the behaviour of the approximate normalizing constants $\hat{p}_L^*(y)$, $p_L(y)$ and the exact normalizing constant $p(y)$, as the sample size increases (see Davison *et al.*, 2006 for a similar argument in a frequentist likelihood setting).

As an illustration, consider data $y = (y_1, \ldots, y_n)$, with $n = 2, \ldots, 35$, drawn from the gamma distribution $\Gamma(e^a, e^b)$, with prior $(a, b) \sim N(0, 10)^2$. The exact posterior normalizing constant $p(y)$ is computed with a bivariate quadrature rule, which is implemented in the function `adapt` of the R package `fCopulae` (Wuertz *et al.*, 2013). Let c_1 and c_2 be positive values and suppose that

$$p(y) = \hat{p}_L^*(y)(1 + b_1 n^{-c_1}) + o(n^{-c_1}) \quad \text{and} \quad p(y) = \hat{p}_L(y)(1 + b_1 n^{-c_2}) + o(n^{-c_2}),$$

for $n \to \infty$. If this is true, then both the quantities $\log\{\hat{p}_L^*(y)/p(y)\}$ and $\log\{p_L(y)/p(y)\}$ would converge to zero as n increases. In addition, if HOA-Laplace is more accurate than the Laplace approximation, then $\log\{\hat{p}_L^*(y)/p(y)\}$ would tend to zero at a faster rate than $\log\{p_L(y)/p(y)\}$. The plot of $\log\{\hat{p}_L^*(y)/p(y)\}$ and $\log\{p_L(y)/p(y)\}$ against $\log n$, in the case the gamma model, and shown on the left of Figure 3.9 gives a firm confirmation

of this. Indeed, the HOA-Laplace approximation converges to zero almost immediately. For larger sample sizes (not shown here) the two methods are indistinguishable.

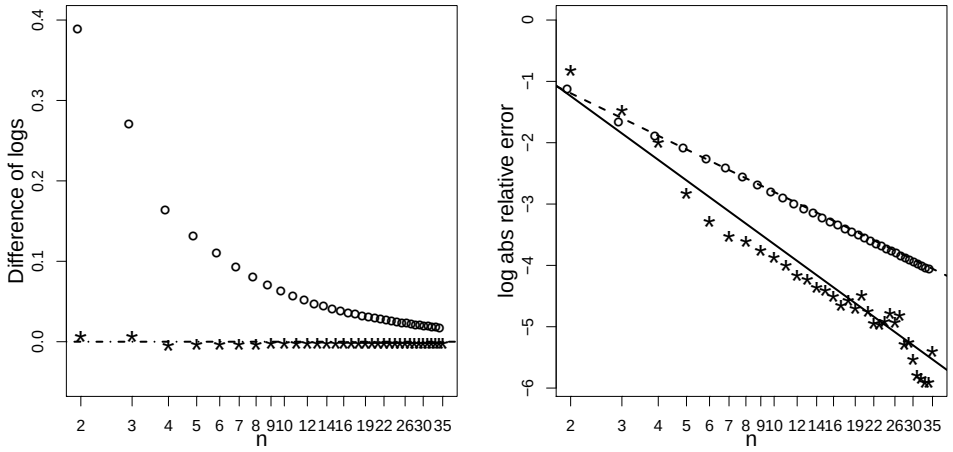


FIGURE 3.9: Study of the asymptotic error of the HOA-Laplace method (——, slope -3/2; - - -, slope -1 and -·-·-, slope 0). Left: log-odds of $\hat{p}^*_L(y)$ $(*)$ and log-odds of $p_L(y)$ $(\circ)$ against $\log n$. Right: log of absolute relative difference between $\hat{p}^*_L(y)$ and $p(y)$ $(*)$ and log of absolute relative difference between $p_L(y)$ and $p(y)$ $(\circ)$.

Moreover, if HOA-Laplace is third-order accurate, then the log-log plot of $|\hat{p}^*_L(y)/p(y)-1|$ against n should be linear with intercept $\log|b_1|$ and slope $-3/2$ ($c_1 = 3/2$). On the other hand the log-log plot of $|p_L(y)/p(y) - 1|$ against n, should be linear with slope $\log|b_2|$ and intercept -1 ($c_2 = 1$). As shown by the plot on the right of Figure 3.9, the two lines (vertically shifted for improving readability) have indeed the claimed intercept.

3.4.2 Examples

To illustrate the accuracy of the HOA-Laplace approximation we consider three examples. The first is a multivariate skew-t distribution (Jones, 2002) with various degree of skewness and with various dimensions. The second example is a probit regression and the third is a nonlinear regression, with 10 parameters. For comparison purposes HOA-Laplace is compared with other typical methods.

Skew-t density

Consider a toy example, where the aim is to compute the normalizing constant of a particular class of skew-t distributions, where the normalizing constant is known. The example was studied by Nott *et al.* (2009) in order to test their improved Laplace approximation method for computing posterior normalizing constants.

The multivariate skew-t density we consider, which is proposed by Jones (2002), is constructed by taking a multivariate t-student density, with mean 0 and scale matrix equal to the identity matrix I, and replacing the marginal distribution for the first component with a suitable skew t distribution. As in Jones (2002, p. 95), we consider the parametrization with parameters a, c and ν, where a and $c > 0$ determine the distribution of the skewed marginal for the first component. The case with $a = c = \nu/2$ leads to the ordinary multivariate t-student distribution with identity scale matrix I and ν degrees of freedom (d.f.). The parameter ν controls the tail behaviour of the distribution.

As an example we consider two values of ν ($\nu = 3$, $\nu = 10$), and consider values of a and c corresponding to zero, moderate and extreme skewness, in 2, 5 and 10 dimensions.

		Skewness			
Two dimensions		None	Minimal	Moderate	Extreme
		$a = c = 1.5$	$a = 4$, $c = 2.5$	$a = 4$, $c = 2$	$a = 4$, $c = 1$
3 d.f.	Laplace	0.6	0.646	0.619	0.493
	HOA-Laplace	1.000	1.000	1.000	1.000
		$a = c = 5$	$a = 13$, $c = 9$	$a = 13$, $c = 6$	$a = 13$, $c = 6$
10 d.f.	Laplace	0.833	0.859	0.832	0.740
	HOA-Laplace	1.000	1.000	1.000	1.000
Five dimensions					
		$a = c = 1.5$	$a = 4$, $c = 2.5$	$a = 4$, $c = 2$	$a = 4$, $c = 1$
3 d.f.	Laplace	0.211	0.211	0.193	0.126
	HOA-Laplace	1.000	1.000	1.000	1.000
		$a = c = 5$	$a = 13$, $c = 9$	$a = 13$, $c = 6$	$a = 13$, $c = 6$
10 d.f.	Laplace	0.506	0.504	0.450	0.320
	HOA-Laplace	1.000	1.000	1.000	1.000
Ten dimensions					
		$a = c = 1.5$	$a = 4$, $c = 2.5$	$a = 4$, $c = 2$	$a = 4$, $c = 1$
3 d.f.	Laplace	0.027	0.027	0.024	0.013
	HOA-Laplace	1.000	1.000	1.000	1.000
		$a = c = 5$	$a = 13$, $c = 9$	$a = 13$, $c = 6$	$a = 13$, $c = 6$
10 d.f.	Laplace	0.172	0.140	0.099	0.056
	HOA-Laplace	1.000	1.000	1.000	1.000

TABLE 3.11: HOA-Laplace and Laplace approximation of the normalizing constant of the multivariate skew t densities in $2, 5$ and 10 dimensions, with 3 and 10 degrees of freedom and zero, minimal, moderate and extreme skewness. The true value is equal to 1.

The results are given in Table 3.11. The required Hessians are computed numerically with the `numDeriv` package (Gilbert & Varadhan, 2012), whereas the integrals are performed with the function `integrate` (R Core Team, 2013). We see that HOA-Laplace approximation gives almost exact results, for all the cases considered (compare these results with those of Nott *et al.*, 2009, p. 1399).

Probit regression

For the second example we consider the binary probit regression applied to Nodal Involvement data used also by Chib (1995), for illustrating his method for computing marginal likelihoods. This dataset refers to a sample of 53 patients with cancer of the prostate and it includes a binary response that takes the value 1 if cancer had spread to the surrounding lumph nodes, and value zero otherwise. The objective is to explain the binary response with five variables: age of patients in years at diagnosis (x_1), level of serum acid phosphate (x_2) considered in logarithmic scale, the result of an X-ray examination, coded 0 if small and 1 if large (x_4), and the pathological grade of the tumor, coded 0 if less serious and 1 if more serious (x_5). We focus on the posterior normalizing constant for three possible models, which correspond to the three models with the highest marginal posterior probability considered in Chib (1995, Tab. 2). As in Chib (1995), for the regression coefficients we assume normal independent priors with mean 0.75 and variance 25.

We compare the HOA-Laplace method with Chib's estimator, which in this case is very easy to compute, and with the usual Laplace approximation. For each model, Chib's estimator is computed from 10^5 runs of the Gibbs sampler (Albert & Chib, 1993), with 10^3 initial values discarded.

Terms fitted	Chib	HOA-Laplace	Laplace
$C + \log x_2 + x_4$	-36.1280	-36.1291	-36.1442
$C + \log x_2 + x_3 + x_4$	-34.5489	-34.5481	-34.5830
$C + \log x_2 + x_3 + x_4 + x_5$	-36.2351	-36.2353	-36.2897

TABLE 3.12: Probit regression with Nodal Involvement Data. Comparison of HOA-Laplace approximation with Chib's and Laplace's approximation for marginal likelihoods.

From the results shown in Table 3.12, and compared to Chib's estimates, we deduce that the Laplace approximation is accurate only to the first decimal place, while the HOA-Laplace method gives results which are accurate to two or three decimal places. However, from a practical point of view the three methods preforms quite similarly.

Nonlinear regression

The last example is a nonlinear regression model with 10 parameters, analyzed also by DiCiccio *et al.* (1997). For this model, we consider the Lubricant data (Bates & Watts, 1988, p. 275), which concern the kinematic viscosity of a lubrificant as a function of

temperature (x_1), and pressure (x_2). The model considered is

$$y_i = \frac{\theta_1}{\theta_2 + x_{1,i}} + \theta_3 x_{2,i} + \theta_4 x_{2,i}^2 + \theta_5 x_{2,i}^3 + (\theta_6 + \theta_7 x_{2,i}^2) x_{2,i} \exp\left\{ -\frac{x_{1,i}}{\theta_8 + \theta_9 x_{2,i}^2} \right\} + \epsilon_i,$$

where the ϵ_i's are independent $N(0, \sigma^2)$ errors, y_i is the response and $x_{j,i}$ denote the $i$the element of the jth covariate, $j = 1, 2$ and $i = 1, \ldots, n$.

As in DiCiccio *et al.* (1997), for the complete parameter $\theta = (\theta_1, \ldots, \log \sigma)$ we adopt independent normal priors centred at the MLE, with standard deviations equal to $n^{1/2}$ times the standard error of the MLE of the parameters.

In this example the required Hessians are computed analytically. We compare the proposed HOA-Laplace method with the usual Laplace approximation (2.8), the Bartlett-corrected Laplace approximation (2.17) (see also DiCiccio *et al.*, 1997), obtained with a large MCMC sample from the posterior, the importance sampling approximation (2.21), with a t-student importance density with 3 d.f., centred at the posterior mode and with scale matrix $\Sigma(\tilde{\theta})$, with the diagonal components of $\Sigma(\tilde{\theta})$ scaled by 1.2. Finally, we consider also numerical integration with modified Gauss-Hermite quadrature rules, as implemented in the R package `bayespack` (Genz & Bornkamp, 2011). The numerical integration is quite expensive here due to the dimensionality of the posterior. Other methods such as (2.19) and the harmonic mean estimator produced much too variable estimates, whereas (2.20) performed slightly worse than importance sampling and are not reported here.

Quadrature	HOA-Laplace	Laplace	Bartlett	Importance
64.0923	64.0841	63.5968	64.1037	64.0890

TABLE 3.13: Nonlinear regression with Lubricant data. Comparison of the HOA-Laplace method with a numerical integration via modified Gauss-Hermite quadrature rule, the Laplace, the Bartlett-corrected Laplace approximation and the importance sampling.

If we consider the quadrature and the importance sampling results as the closest to the truth, then the HOA-Laplace approximation is the most accurate, despite the dimension being as high as 10.

3.4.3 Remarks

By combining Chib's idea with the Laplace approximation for marginal posterior distributions (Tierney & Kadane, 1986), we obtain a higher-order Laplace approximation (HOA-Laplace) for posterior normalizing constants. We show, both theoretically and

empirically, that HOA-Laplace has third-order accuracy. Moreover, in the examples considered in this section HOA-Laplace is competitive to other state-of-art methods.

An advantage of the HOA-Laplce approximation is that it requires only nested optimizations, and evaluations of the log-posterior Hessian as well as univariate numerical integrations. In general, these quantities can be obtained from any software that performs numerical calculations, such as R . Hence the method is analytic and does not require simulations from the posterior density.

Chapter 4

Contributions on Likelihood-free Methods

The summary of the data on a given model offered by the likelihood function is the basis 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 models with many latent variables.

As outlined in Section 2.4.1, and by the contribution in Section 3.2, it is possible to deal with such complex models by means of suitable pseudo-likelihood functions. On the other hand, ABC methods are also a valid alternative in these contexts (see Sect. 2.4.2), as they bypass the computation of the full likelihood by simulating from the full model.

In this chapter we present two original contributions to the ABC literature, which combine pseudo-likelihood functions with ABC. In particular, the first contribution, presented in Section 4.1, aims at finding good summary statistics by combining ABC with composite likelihoods. The second, discussed in Section 4.2, shows how to automatically build a proposal distribution for ABC algorithms with an MCMC or IS step, by using the theory of quasi-likelihoods (McCullagh, 1991).

4.1 Approximate Bayesian computations with composite score functions

When the full likelihood function is intractable, it is possible to resort to pseudo-likelihood functions, which are intended as surrogates of the full likelihood. An important class of such pseudo-likelihoods is given by *composite likelihoods* (see Sec. 2.4.1). 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).

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 Sec. 2.4.2). In most applications, the probability of an exact match of the simulated data with the observed data is negligible, or zero. 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 true posterior distribution as the distance tends to zero, provided 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 demanding.

We show that composite likelihoods and ABC can be fruitfully integrated in order to obtain accurate approximations to the posterior distribution of the parameter of interest, without having to specify *ad hoc* summary statistics. 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 in a full likelihood, 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 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 full 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 overly concentrated posterior distributions (Sec. 2.4.1), the proposed ABC-cs is proved to automatically give correctly adjusted posterior approximations.

There have been other attempts to merge composite likelihoods with the ABC framework. For instance, Erhardt & Smith (2012), in the context of spatial extremes, combine composite likelihoods with ABC and show that this approach tends to work better than other existing methods. Mengersen *et al.* (2013) use the composite score function with the empirical likelihood 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 (2011, 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 the ABC-cs method in some sense also relies on an auxiliary model likelihood, that is, the composite likelihood. As happens in indirect inference, the closer the auxiliary model is to the true model, the more accurate the parameter estimates will be. However, the ABC-cs approach is less computationally demanding than indirect inference methods 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).

4.1.1 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 (see Sect. 2.5.2), 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 the composite score function $c\ell_\theta(\theta; y)$ (see Sect. 2.5.1 for discussion on composite likelihoods), evaluated at $\hat{\theta}_c$ computed from the observed data y^{obs}, as a summary of the data. We justify this choice by starting from a full

computable likelihood function, and then we extend the proposed ABC-cs algorithm to general likelihood functions.

ABC with score functions

Let us assume that the model belongs to a full exponential family with density $p(y;\varphi) = h(y)\exp\{\varphi^T s(y) - k(\varphi)\}$, where $h(y) > 0$, φ 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). On the other hand, let us consider the score function $\ell_\varphi(\varphi_0; y) = \partial\ell(\varphi; y)/\partial\varphi|_{\varphi=\varphi_0} = s(y) - \partial k(\varphi)/\partial\varphi|_{\varphi=\varphi_0}$ as a summary statistic in ABC. Considering as distance any norm-p, the distance among the scores is exactly the distance among the sufficient statistics, regardless of the fixed value φ_0. Therefore, the ABC posterior with the score function is exact for $\epsilon \to 0$. When the model is reparametrized using $\theta = \theta(\varphi)$, the ABC algorithm would be obviously still based on the sufficient statistic $s(y)$.

Consider now a generic model $p(y;\theta)$. 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 maximum likelihood estimate and a is an ancillary statistic, which means that its distribution does not depend on θ. Hence, we could replace $p(y;\theta)$ with $p(\hat{\theta}, a; \theta)$, and the latter can be factorized as

$$p(\hat{\theta}, a; \theta) = p(\hat{\theta}|a; \theta)p(a)\,.$$

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

$$\ell^{\text{TE}}(\theta; y) = \ell(\theta; y^{\text{obs}}) - \ell(\hat{\theta}^{\text{obs}}; y^{\text{obs}}) + \{\varphi(\theta) - \varphi(\hat{\theta}^{\text{obs}})\}^T s(y)\,, \tag{4.1}$$

where $\hat{\theta}^{\text{obs}}$ is the maximum likelihood estimate at the observed data point y^{obs}, $s(y) = \partial\ell(\theta; y)/\partial\theta|_{\theta=\hat{\theta}^{\text{obs}}} = \ell_\theta(\hat{\theta}^{\text{obs}}; y)$, and $\varphi(\theta) = \varphi(\theta; y^{\text{obs}})$ is a one-to-one reparameterization dependent on the observed data y^{obs} (see 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 φ. 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.

As in a full exponential family, the best summary statistic for ABC with the tangent exponential model (4.1) would be the sufficient statistic $s(y)$, which is the score function of the original model evaluated at $\hat{\theta}^{\text{obs}}$. Note that $s(y^{\text{obs}}) = 0$. This motivates the use of the score function evaluated at $\hat{\theta}^{\text{obs}}$ as an approximate optimal summary statistic in ABC for a general model.

Example. *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 $\ell_\theta(\theta; y) = -\theta^{-2}\sum_{i=1}^n y_i + \theta^{-3}\sum_{i=1}^n y_i^2 - n/\theta$, which implies that $\hat{\theta}$ is the positive solution of a quadratic equation. The sufficient statistic for the tangent exponential model has the same dimension as the parameter and is given by $s(y) = \ell_\theta(\hat{\theta}^{\text{obs}}; y)$.

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 apply the ABC algorithm **3** (Sec. 2.4.2), with distance $\rho(v, w) = ||v - w||_1$ and with summary statistics given respectively by $t(y)$, $s(y)$, and 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 ϵ as the quantile of level 0.1% of the observed distances, thus accepting 10^4 values. These ϵ values are respectively 31.264, 0.052 and 0.237.

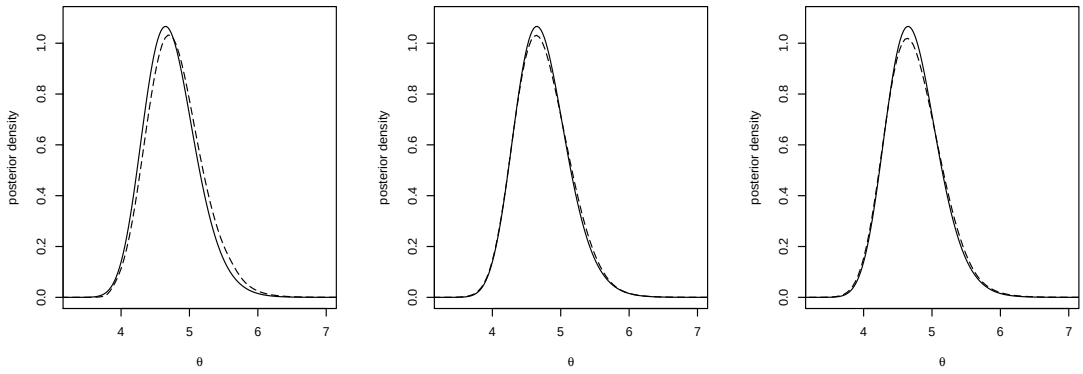


FIGURE 4.1: Normal parabola. In all panels the solid line corresponds to the exact posterior, while the dashed lines correspond to ABC approximations using $t(y)$ (left panel), $t_1(y)$ (central panel), $s(y)$ (right panel).

Figure 4.1 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 ϵ (31.264). Only three of the 10^7 proposed values of θ 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 $s(y)$, which is not sufficient for this model, gives an approximation to the posterior with accuracy comparable with ABC with the minimal sufficient statistic $t_1(y)$.

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 the previous example 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 $s(y)$ is invariant with respect to such transformations.

An apparent drawback of the ABC algorithm with $s(y)$ is the dependence on the parameterization θ. However, one-to-one reparameterizations of the model only rescale the summary statistic $s(y)$. Indeed, let $\omega = \omega(\theta)$ be a reparameterization with corresponding log-likelihood $\bar{\ell}(\omega; y) = \ell(\theta(\omega); y)$, and score function $\bar{\ell}_\omega(\omega; y) = \{\partial\theta(\omega)/\partial\omega\}\ell_\theta(\theta(\omega); y)$. The summary statistic becomes $\bar{s}(y) = \bar{\ell}_\omega(\hat{\omega}^{\text{obs}}; y) = \{\partial\theta(\omega)/\partial\omega\}|_{\omega=\hat{\omega}^{\text{obs}}}\, s(y)$. Let us assume that $\rho(v, w) = ||v - w||_1$, although the following result will hold for any p-norm. Then, by standard properties of norms (see Noble & Daniel, 1988, Sect. 5.3), we have that

$$\begin{aligned} ||\bar{s}(y^{\text{obs}}) - \bar{s}(y)||_1 &= ||\{\partial\theta(\omega)/\partial\omega\}|_{\omega=\hat{\omega}^{\text{obs}}}\{s(y^{\text{obs}}) - s(y)\}||_1 \\ &\leq k\ ||s(y^{\text{obs}}) - s(y)||_1\,, \end{aligned}$$

where k is a suitable positive constant depending on $\hat{\omega}^{\text{obs}}$. In view of this, $||s(y^{\text{obs}}) - s(y)||_1 \leq \epsilon$ implies that $||\bar{s}(y^{\text{obs}}) - \bar{s}(y)||_1 \leq \epsilon^* = k\,\epsilon$. Therefore, $\epsilon \to 0$ implies $\epsilon^* \to 0$ as well. This shows that the validity of the ABC algorithm with the score function is invarinat to reparameterizations.

Despite the good properties of ABC with the score function, unfortunately in typical applications of the ABC method the likelihood function, as well the score function, are of course unavailable. This motivates the extension to composite likelihoods.

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 was seen in the previous

section for a full likelihood, we propose the composite score function as a summary statistic in ABC. This defines an algorithm, called ABC-cs. In terms of the ABC accept-reject Algorithm 3, ABC-cs replaces the matching condition

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

with

$$\rho\{c\ell_\theta(\hat{\theta}_c^{\text{obs}}; y^{\text{obs}}), c\ell_\theta(\hat{\theta}_c^{\text{obs}}; y)\} \leq \epsilon \ ,$$

where $\hat{\theta}_c^{\text{obs}}$ is the MCLE computed from y^{obs}. This choice is computationally convenient since $c\ell_\theta(\hat{\theta}_c^{\text{obs}}; y^{\text{obs}}) = 0$ and we only need to evaluate $c\ell_\theta(\hat{\theta}_c^{\text{obs}}; y)$.

An advantage of ABC-cs is that the composite score statistic has the same dimension as θ, so 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.

The proposed ABC-cs algorithm gives a valid approximation to the posterior distribution even if the composite score function does not satisfies the information identity, as a full score function. In order to recover the information identity, the rescaled composite score function (see, *e.g.*, Pace & Salvan, 1997, Chap. 4)

$$g(\theta; y) = K(\theta)J(\theta)^{-1}c\ell_\theta(\theta; y) = A(\theta)c\ell_\theta(\theta; y)$$

should be considered, where recall that $K(\theta) = E_\theta\{-\partial c\ell_\theta(\theta; y)/\partial\theta^T\}$ and $J(\theta) = \text{var}_\theta\{c\ell_\theta(\theta; y)\}$. 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

$$\begin{aligned} K_g(\theta) &= E_\theta\left\{-\frac{\partial}{\partial\theta^T}g(\theta; Y)\right\} \\ &= -\left\{\frac{\partial}{\partial\theta^T}A(\theta)\right\}E_\theta\{c\ell_\theta(\theta; Y)\} - A(\theta)E_\theta\left\{\frac{\partial}{\partial\theta^T}c\ell_\theta(\theta; Y)\right\} = G(\theta)\,. \end{aligned}$$

Since $K_g(\theta) = J_g(\theta) = G(\theta)$, the rescaled composite score $g(\theta; y)$ satisfies the information identity as the full score function. Moreover, since $A(\theta) \neq 0$ and $A(\theta)^{-1}$ is finite, the estimating equation $g(\theta; y) = 0$ gives the same estimator $\hat{\theta}_c$ of $c\ell_\theta(\theta; y) = 0$. The use of $g(\hat{\theta}_c^{\text{obs}}; y)$ as a summary statistic for ABC leads to an approximate posterior with the correct curvature (see Pauli *et al.*, 2011). Nevertheless, this rescaling turns out to be irrelevant in the ABC-cs algorithm. Indeed, with an argument similar to the one used

for the invariance to reparameterizations of ABC with the score function in the previous section, we have that

$$\begin{aligned} ||g(\hat{\theta}_c^{\text{obs}}; y^{\text{obs}}) - g(\hat{\theta}_c^{\text{obs}}; y)||_1 &= ||A(\hat{\theta}_c^{\text{obs}})\{c\ell_\theta(\hat{\theta}_c^{\text{obs}}; y^{\text{obs}}) - c\ell_\theta(\hat{\theta}_c^{\text{obs}}; y)\}||_1 \\ &\leq h\ ||c\ell_\theta(\hat{\theta}_c^{\text{obs}}; y^{\text{obs}}) - c\ell_\theta(\hat{\theta}_c^{\text{obs}}; y)||_1\,, \end{aligned}$$

where h is a suitable positive constant. Again, $||c\ell_\theta(\hat{\theta}_c^{\text{obs}}; y^{\text{obs}}) - c\ell_\theta(\hat{\theta}_c^{\text{obs}}; y)||_1 \to 0$ implies that $||g(\hat{\theta}_c^{\text{obs}}; y^{\text{obs}}) - g(\hat{\theta}_c^{\text{obs}}; y)||_1 \to 0$. Therefore, ABC with $c\ell_\theta(\hat{\theta}_c^{\text{obs}}; y)$ gives an approximation to the posterior distribution with the correct curvature, without requiring the rescaling factor $K(\theta)J(\theta)^{-1}$, which can be cumbersome to evaluate. On the other hand, the use of the calibrated composite likelihood $cL_c(\theta)$ to obtain the posterior distribution (2.32) requires the computation of the adjustment factor $1/\bar{\omega}$, which explicitly needs the evaluation of $J(\theta)$ and $K(\theta)$.

4.1.2 Examples

In the examples discussed in this section we use composite marginal likelihood functions, although different model structures might lead to different choices of suitable composite likelihoods.

Whenever possible, ABC and ABC-cs posteriors are compared also with the "exact" Bayesian posterior, *e.g.* the posterior based on the full likelihood, possibly approximated using MCMC methods. The distance used in all examples is the absolute norm.

Equi-correlated normal model

This example, considered in Cox & Reid (2004) and in Pauli *et al.* (2011) among others, focuses on Bayesian inference based on the pairwise log-likelihood (2.28) for the correlation coefficient ρ of an equi-correlated multivariate normal distribution.

Let Y_i be independent realizations of a q-variate normal random variable with standard margins, and let $\text{cor}(Y_{ir}, Y_{is}) = \rho$, for $r, s = 1, \ldots, q$, $r \neq s$ $(i = 1, \ldots, n)$, with $\rho \in (-1/(q-1), 1)$. The pairwise log-likelihood (2.28) is

$$p\ell(\rho; y) = -\frac{nq(q-1)}{4}\log(1-\rho^2) - \frac{q-1+\rho}{2(1-\rho^2)}SS_W - \frac{(q-1)(1-\rho)}{2(1-\rho^2)}\frac{SS_B}{q}, \qquad (4.2)$$

where $SS_W = \sum_{i=1}^n \sum_{r=1}^q (y_{ir} - \bar{y}_i)^2$, $SS_B = q^2 \sum_{i=1}^n \bar{y}_i^2$, $\bar{y}_i = \sum_{r=1}^q y_{ir}/q$, and the associated score function is

$$p\ell_\rho(\rho; y) = \frac{nq(q-1)\rho}{2(1-\rho^2)} - \frac{1+\rho^2+2(q-1)\rho}{2(1-\rho^2)^2}SS_W + \frac{(q-1)(1-\rho)^2}{2(1-\rho^2)^2}\frac{SS_B}{q}\ . \qquad (4.3)$$

We reparameterize in terms of $\theta = \text{logit}\{(\rho(q-1)+1)/q\}$, *i.e.* the logistic transformation, and for θ we assume a $N(0, 5)$ prior.

The ABC-cs uses the pairwise score function (4.3) evaluated at the maximum pairwise likelihood estimate (MPLE), whereas the usual ABC algorithm is implemented using the two-dimensional sufficient statistic (SS_B, SS_W). As an example, a sample of $n = 50$ is drawn from the model with $q = 50$ and $\rho = 0.5$. Both algorithms are run for 10^6 proposal values from the prior and ϵ is fixed to the 0.1% quantile of the absolute distances between the statistics. Results are compared also with the pairwise posterior

$$\pi_{pl}(\theta|y) \propto \pi(\theta) \exp\{p\ell(\theta; y)\} \ , \tag{4.4}$$

and with the pairwise posterior (2.32) based on the calibrated pairwise likelihood.

The left panel of Figure 4.2 compares the ABC-cs posterior for θ, with the exact, the pairwise (4.4) and the calibrated pairwise (2.32) posteriors. The right panel of Figure 4.2 compares ABC-cs and ABC with the full posterior.

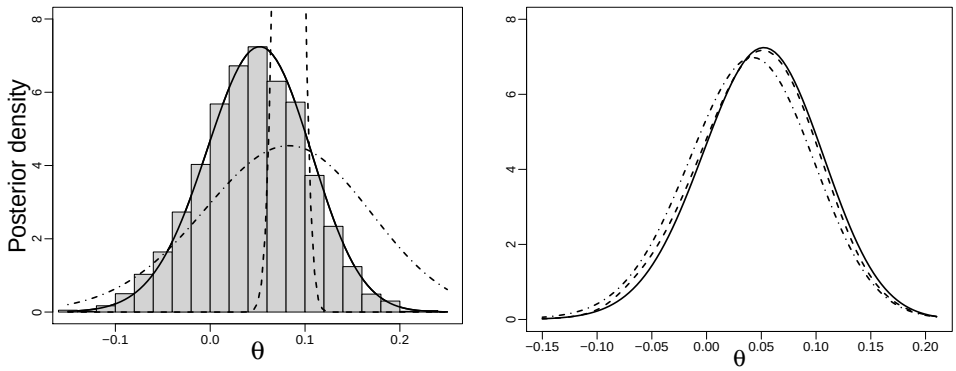


FIGURE 4.2: Equi-correlated normal model. (Left) ABC-cs posterior (histogram), compared with $\pi(\theta|y)$ (continuous line), $\pi_{pl}(\theta|y)$ (dashed) and $\pi^c_{pl}(\theta|y)$ (dot-dashed). (Right) ABC-cs posterior (dashed) compared with the ABC (dot-dashed) and the full posterior $\pi(\theta|y)$ (continuous).

Figure 4.2 highlights several interesting features. As is well known, the posterior (4.4) can be wrongly too concentrated (see also Pauli *et al.*, 2011; Smith & Stephenson, 2009; Ribatet *et al.*, 2012), whereas the calibrated pairwise posterior (2.32) may be the opposite. On the other hand, the ABC-cs posterior follows the full posterior very closely. The ABC posterior based on the sufficient statistics is slightly worse than ABC-cs. This may be due to the particular form of sufficient statistic used in the ABC algorithm (see Section 3.1).

We now assume that the model has mean vector μ and covariance matrix $\Sigma_{rs} = \rho\sigma^2$, for $r \neq s$ and $\Sigma_{rr} = \sigma^2$ $(r, s = 1, \ldots, q)$. In this case $\hat{\theta}_c$ is fully 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 is given in (3.8). We assume the components of the parameter $\theta = (\mu, \tau, \kappa)$, with $\tau = \log \sigma^2$ and $\kappa = \text{logit}\{(\rho(q-1)+1)/q\}$, a priori independent with $N(0, 5)$ marginal distributions.

A sample with $n = 50$ is drawn from the model with $q = 50$, $\mu = 0$, $\sigma^2 = 1$ and $\rho = 0.5$. For ABC, the summary statistic is the sufficient statistic, while for ABC-cs the summary is the pairwise score function evaluated at the MPLE. The simulation from the ABC and ABC-cs posterior is obtained by importance sampling. The importance function is a t-student density with 5 degrees of freedom centred at the maximum likelihood estimate (MLE), with scale matrix equal to 3 times the inverse of the negative log-posterior Hessian. We consider 10^3 final samples obtained after fixing ϵ to the 0.1% quantile of the observed distances.

The various marginal posterior approximations are shown in Figure 4.3 by means of box-plots. Also in this case, the non-calibrated pairwise posterior is too narrow,

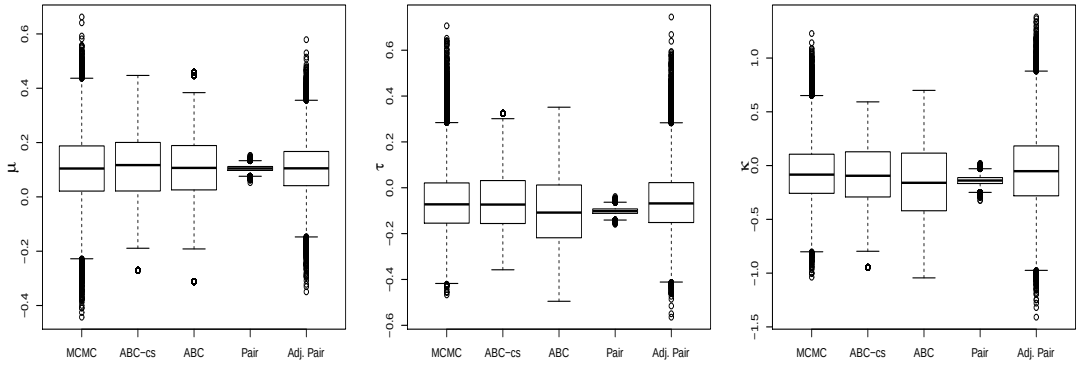
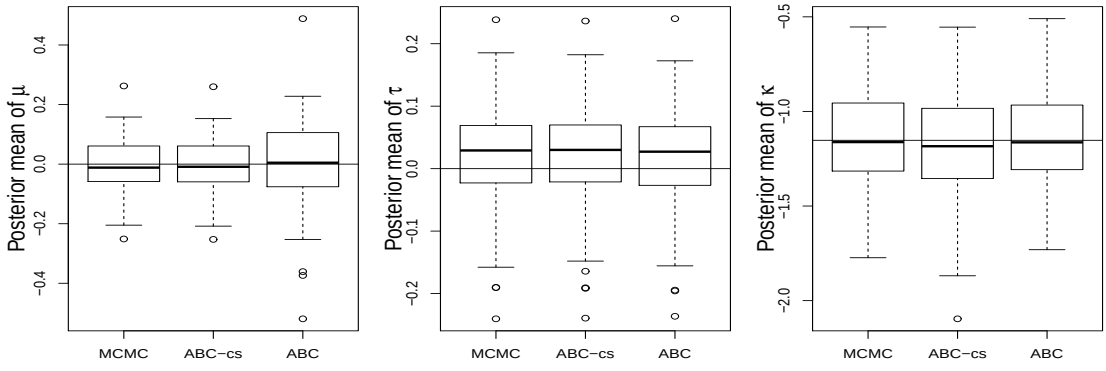


FIGURE 4.3: Equi-correlated normal model (continued). ABC-cs posterior compared with the full (MCMC), the pairwise (Pair), the calibrated pairwise (Adj. Pair) and the ABC posterior.

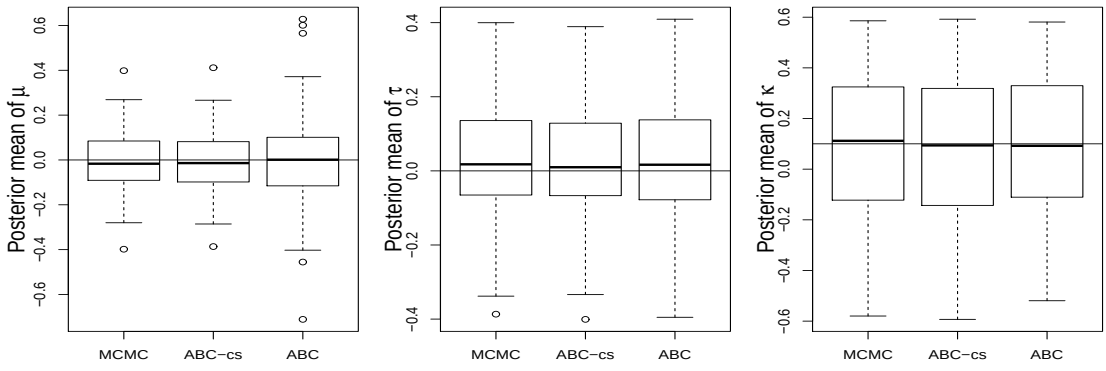
whereas the calibrated pairwise posterior, the ABC-cs and ABC are all quite similar to the full posterior (MCMC), approximated via a random walk Metropolis-Hastings algorithm. This is not surprising, since the model is a full exponential family of order three and ABC uses exactly the sufficient statistic 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 Sect. 4.1.1).

We also compare the posterior mean of the ABC and ABC-cs posteriors in a simulation study over 100 Monte Carlo trials. The data are generated from the model with $\mu = 0$, $\sigma^2 = 1$, $\rho = (0.2, 0.5, 0.9)$, $n = 30$, and $q = 20$. At each simulated dataset, the ABC, the ABC-cs and the exact posteriors are approximated as in the example above. From the

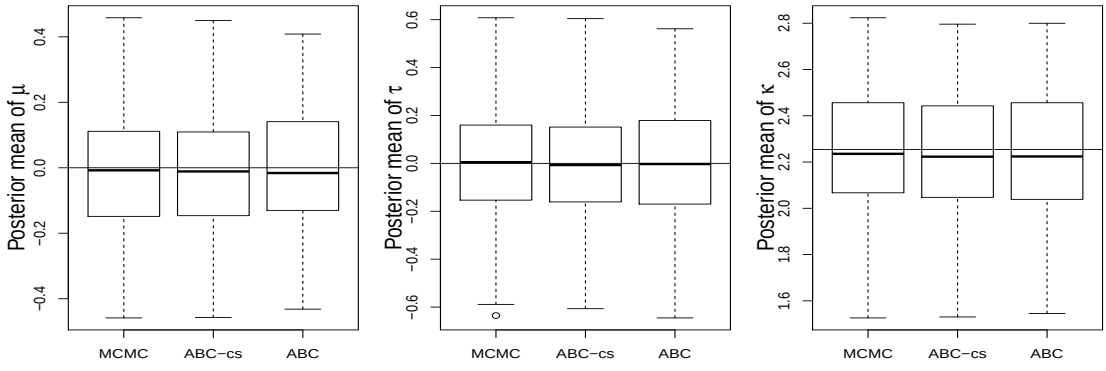
simulations (see Figure 4.4), we notice that ABC and ABC-cs posterior means perform quite similarly to the full posterior mean, as expected from the comments above.



(a) $\rho = 0.2$ ($\kappa \approx -1.15$)



(b) $\rho = 0.5$ ($\kappa \approx 0.1$)



(c) $\rho = 0.9$ ($\kappa \approx 2.25$)

FIGURE 4.4: Equi-correlated normal model (continued). Simulation study based on 100 Monte Carlo trials, with $\mu = 0$, $\sigma = 1$ ($\tau = 0$).

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. Maximum likelihood

analysis in this contexts may be difficult because it involves multivariate integrals whose dimension equals the cluster sizes.

Let us focus on a multilevel probit model with constant cluster sizes. In particular, let S_i be a latent q-variate normal with mean $\gamma_i = X_i\beta/\sigma$, with β a vector of unknown regression coefficient, σ a known scale parameter and X_i the design matrix for unit i, and covariance matrix Σ, with $\Sigma_{hh} = \sigma^2$, $\Sigma_{hk} = \sigma^2\rho$, $h \neq k$ $(i = 1, \ldots, n)$. Then, the observed Y_{ih} is equal to 1 if $S_{ih} > 0$, and 0 otherwise $(h = 1, \ldots, q)$.

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

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

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 ρ, and $\gamma_{ih} = x_{ih}\beta/\sigma$ is the h component of γ_i $(i = 1, \ldots, n, h, k = 1, \ldots, q)$.

As an example, we consider data generated with $\beta_0 = \rho = 0.5$, $\beta_1 = \sigma = 1$, $n = 50$ and $q = 7$, where β_0 is the intercept and β_1 the coefficient of a covariate, which has been generated from $U(-1, 1)$. For the parameter $\theta = (\beta_0, \beta_1, \kappa)$ with $\kappa = \text{logit}((\rho(q - 1) + 1)/q)$ a trivariate normal prior with independent components $N(0, 5)$ is assumed. For ABC we take as summary statistic the counts over individuals at each time point h $(h = 1, \ldots, q)$, as q-dimensional summary statistic. This choice does encounter the curse of dimentsionality as the number of time points q increases. The absolute norm among the statistics for ABC is $\sum_{h=1}^{q} |\sum_{i=1}^{n}(y_{ih} - z_{ih})|$, whereas for ABC-cs we consider the absolute norm of the difference among pairwise scores computed numerically and evaluated at the observed MPLE. We consider 10^3 final samples drawn form the ABC and ABC-cs posteriors after fixing ϵ to the 0.1% quantile of the observed distances. The sampling is done via importance sampling, with a t-student importance density, with 5 degrees of freedom, centred at the MPLE and with scale matrix equal to 13 times the inverse Hessian of negative pairwise log-posterior.

The various marginal posteriors are shown in Figure 4.5. For comparison we report also an expensive MCMC approximation of the posterior based on the full likelihood evaluated by available `fortran` code from Alan Genz's website (`http://www.math.wsu.edu/faculty/genz/homepage`), which for moderate to large values of q tend to be too slow or unstable. We notice that, in this example, occasional likelihood evaluations gave negative values, which within MCMC were treated as unacceptable, and therefore

rejected. Clearly, the ABC posterior is quite different from the target (MCMC), whereas ABC-cs posterior gives a more accurate approximation to the true posterior.

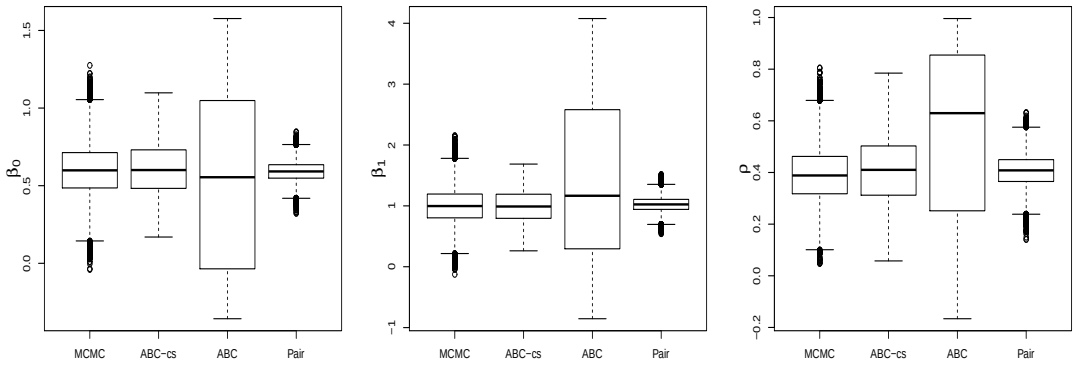


FIGURE 4.5: Correlated binary data. ABC-cs posterior compared with the ABC, the pairwise (Pair), the exact posterior (MCMC) for a simulated dataset with $n = 50$ and $q = 7$.

A simulation study is conducted over 100 Monte Carlo samples, where the covariate x_{ih} are simulated as previously, with $\beta_0 = 0.5$, $\beta_1 = 1$, $n = 50$, $q = 7$ and $\rho = \{0.2, 0.5, 0.9\}$. For each simulated dataset, we consider the mean of 10^3 final samples drawn from ABC and ABC-cs posteriors, respectively, via the importance sampling with $\epsilon = 1\%$. The simulation algorithm is the same as above. Since for some datasets, occasional evaluations of the likelihood gave negative values, results based on the full posterior are not reported.

From the simulations shown in Figure 4.6, it is evident that the ABC mean can perform very poorly, especially for extreme correlation values. On the other hand, the ABC-cs mean is less biased and more precise.

MA(2) process

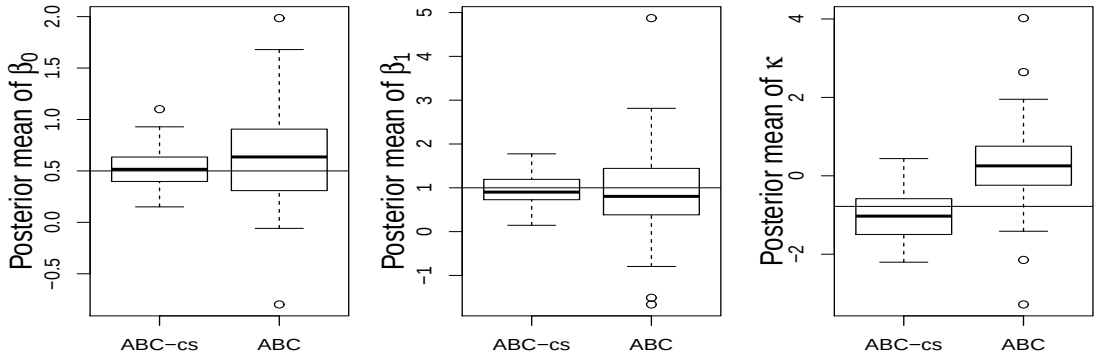
Consider an MA(p) process, defined as

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

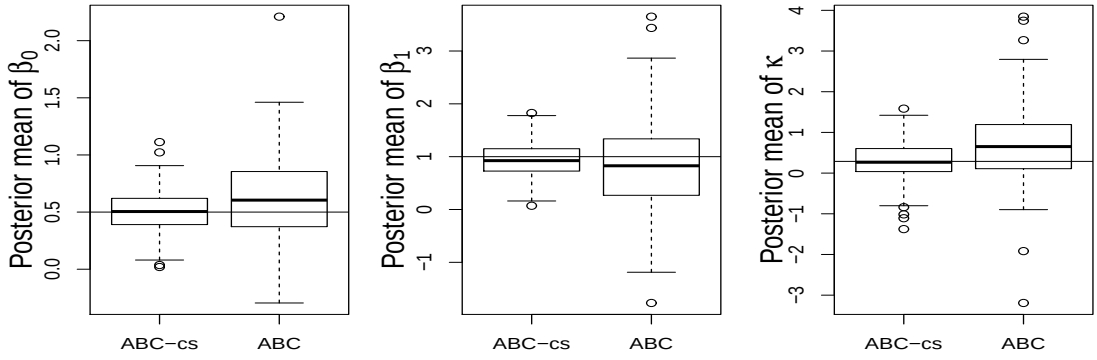
where u_t $(t = 1, \ldots, q)$, is an independent sequence of normals $N(\mu, \sigma^2)$, and θ_i $(i = 1, \ldots, p)$, must satisfy the identifiability conditions, namely that the roots of the polynomial

$$Q(x) = 1 - \sum_{i=1}^{p} \theta_i x^i$$

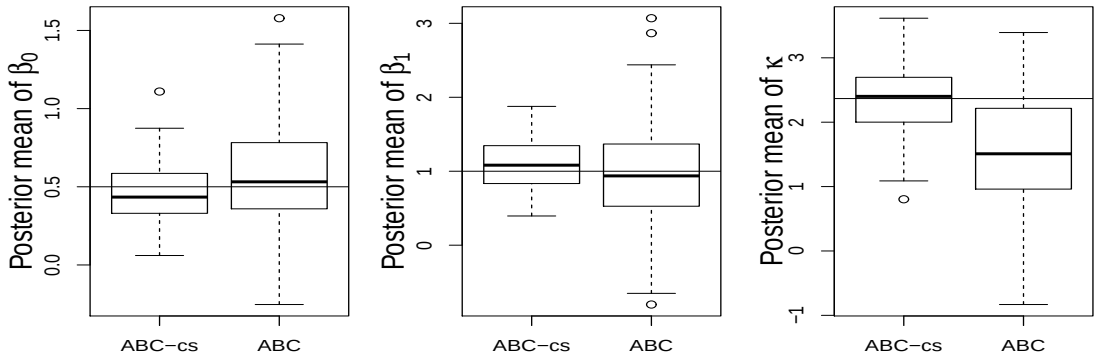
are all outside of the unit circle in the complex plane. This stochastic process is typically used for time series analyses.



(a) $\rho = 0.2$ ($\kappa = -0.78$)



(b) $\rho = 0.5$ ($\kappa = 0.29$)



(c) $\rho = 0.9$ ($\kappa = 2.37$)

FIGURE 4.6: Correlated binary data. Simulations based on 100 Monte Carlo trials, with $\beta_0 = 0.5$, $\beta_1 = 1$.

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 owning 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 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

$$-2 < \theta_1 < 2, \quad \theta_1 + \theta_2 > -1, \quad \theta_1 - \theta_2 < 1\,,$$

and 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 (Hjort & Varin, 2008; Varin, 2008) of the form

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

As in Marin *et al.* (2012), we draw $n = 100$ values from the MA(2) model, with parameters $(\theta_1, \theta_2) = (0.6, 0.2)$. For the ABC-cs posterior the triplewise score is evaluated at the observed MCLE. A sample of 10^3 final values is drawn from the ABC and ABC-cs posteriors. These samples are obtained generating from the prior and ϵ is fixed to the 0.1% quantile of the observed distances. For illustration purposes the ABC and ABC-cs posteriors are compared also with the "exact" posterior approximated with a random walk Metropolis-Hastings algorithm. From the posteriors, shown in Figure 4.7, we notice that ABC approximation tends to be slightly worse than than ABC-cs.

A simulation study is performed, with 100 Monte Carlo samples drawn from the true model with the parameter $(\theta_1, \theta_2) = (0.6, 0.2)$. For each simulated dataset, we run ABC and ABC-cs with 10^3 final samples and ϵ fixed to the 0.1% quantile of the observed distances. Over this final draws the average is taken and it is compared also with the mode of the exact posterior. The simulation results are plotted in Figure 4.8. Both ABC methods give reasonable results when compared to exact posterior, with ABC-cs being overall preferable to ordinary ABC.

4.1.3 Remarks

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 is, or pilot runs or *ad hoc* summaries of the data. The

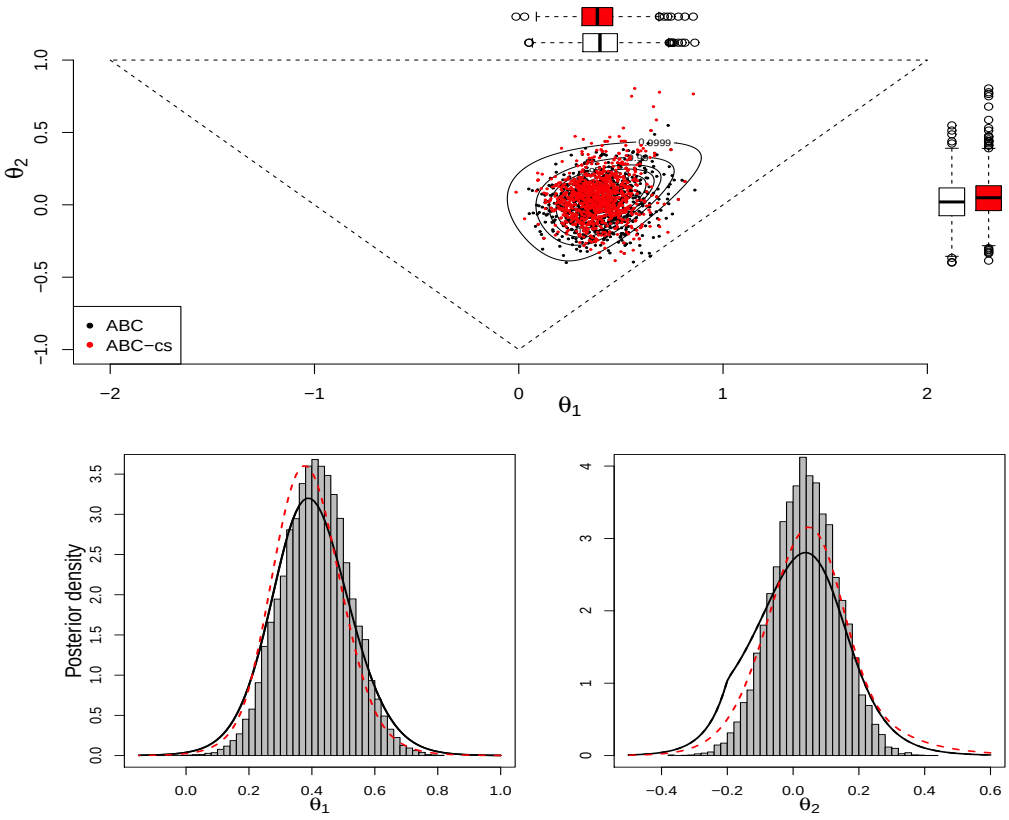


FIGURE 4.7: MA(2) model. Top panel: comparison of the level sets (in black) of the posterior distribution against simulated values with ABC (black dots) and ABC-cs (red dots), with box-plots of the ABC posterior (black) against ABC-cs (red). Bottom-left (bottom-right) panel: histogram of the marginal posterior of θ_1 (θ_2), compared with ABC (continued) and ABC-cs (dashed red coloured).

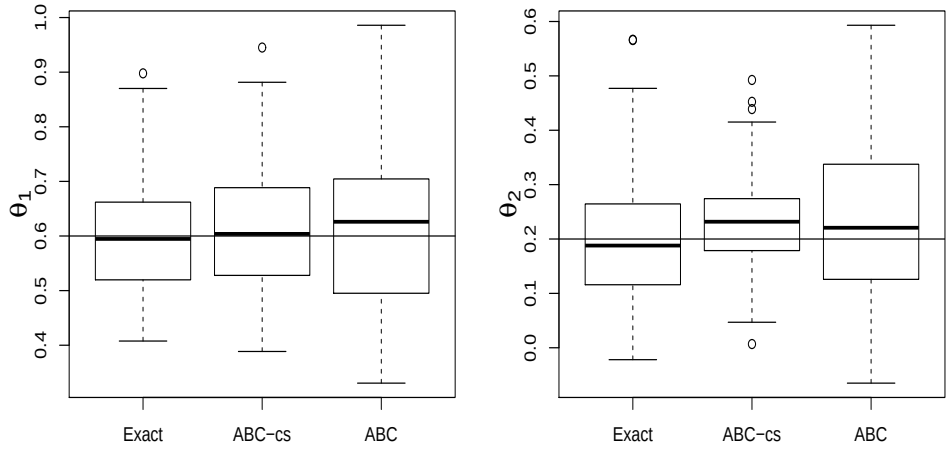


FIGURE 4.8: MA(2) model. Comparisons of the exact posterior mode, ABC and ABC-cs posterior mean in 100 Monte Carlo trials, with $(\theta_1, \theta_2) = (0.6, 0.2)$ (horizontal lines).

proposed approach can be fruitfully used within more elaborate Monte Carlo algorithms, such as MCMC, or sequential Monte Carlo methods, although this possibility has not

been explored yet.

The success of the ABC-cs procedure depends on how good is the composite likelihood as a surrogate for the full model likelihood, 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, 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 & Roncetti, 2009).

4.2 A quasi-likelihood proposal for ABC

ABC is a set of approximation methods useful for Bayesian inference when the likelihood function $L(\theta)$ is analytically or computationally intractable. However, as outlined in Section 2.5.2 (see also Sec. 4.1), ABC has several issues. Among these, we note the choice of the summary statistic, on which an original contribution was shown in Section 4.1. The second issue is related to the inefficiency of the original ABC accept-reject algorithm (see Algorithm 3) when the prior and the bulk of the likelihood are noticeably different. For instance, with flat or improper priors, *i.e.* $\int \pi(\theta)\mathrm{d}\theta = \infty$, Algorithm 3 cannot be used.

Assume that a suitable summary statistic $\eta(\cdot)$ is available, with the same dimension as θ, and consider instances where the original ABC accept-reject algorithm is computationally too expensive because the prior is vague. To simplify notation, let $\eta = \eta(y)$ and $\eta^{\text{obs}} = \eta(y^{\text{obs}})$. At present, there are two ABC algorithms to deal with this: the ABC-MCMC proposed by Marjoram *et al.* (2003) and the ABC-IS algorithm based on IS (see, *e.g.*, Fearnhead & Prangle, 2012). The former bypasses prior simulation by drawing candidate values from a suitable proposal distribution $q(\cdot)$, and then the proposed values are evaluated in a Metropolis-Hastings-like acceptance step. As in the usual MCMC setting, the proposal distribution can be an independent kernel or a Markov kernel. The ABC-MCMC algorithm with an independent proposal distribution $q(\cdot)$ is given in Algorithm 6 (see Algorithm 3 of Marin *et al.*, 2012, for ABC-MCMC with a Markov kernel).

Result: A dependent sample $(\theta^{(1)}, \ldots, \theta^{(m)})$ from $\pi(\theta|\eta^{\text{obs}})$
Data: ϵ, $\eta(\cdot)$, initial value $\theta^{(0)}$ and $q(\cdot)$
for $t = 1 \rightarrow m$ **do**
  repeat
1 draw $\theta^* \sim q(\cdot)$
2 draw $y \sim f(y; \theta^*)$ and set $\eta = \eta(y)$
3 draw $u \sim U(0, 1)$
4 compute $R = \frac{\pi(\theta^*)q(\theta^{(t-1)})}{\pi(\theta^{(t-1)})q(\theta^*)}$
  until $\rho(\eta, \eta^{obs}) \leq \epsilon$ *and* $u \leq R$;
5 set $\theta^{(t)} = \theta^*$
end

Algorithm 6: ABC-MCMC sampler.

Algorithm 6 needs suitable starting values $\theta^{(0)}$, as well as a threshold ϵ. While $\theta^{(0)}$ can be found by an initial run of the original accept-reject algorithm or by trial-and-error,

the threshold can be fixed to some lower quantile of the distances among the summary statistics.

The ABC-IS method, given in Algorithm 7 (see also Fearnhead & Prangle, 2012, for another version), uses the importance density $f(\theta)$ to draw suitable candidate values, which are then weighted according to the prior and the distance among the summary statistics.

Result: A weighted sample $(\theta^{(1)}, \ldots, \theta^{(m)})$ from $\pi(\theta|\eta^{\text{obs}})$
Data: ϵ, $\eta(\cdot)$, and importance density $f(\theta)$
for $t = 1 \rightarrow m$ **do**
  repeat
1 draw $\theta^* \sim f(\cdot)$
2 draw $y \sim f(y; \theta^*)$ and set $\eta = \eta(y)$
3 compute $\omega^* = \frac{\pi(\theta^t)}{f(\theta^{(t)})}$
  until $\rho(\eta, \eta^{obs}) \leq \epsilon$;
4 set $(\theta^{(t)}, \omega^{(t)}) = (\theta^*, \omega^*)$
end

Algorithm 7: ABC-MCMC sampler.

Also Algorithm 7 must deal with ϵ, which can be fixed as for ABC-MCMC.

The crucial point with ABC-MCMC (ABC-IS) is how to choose a good $q(\theta)$ ($f(\theta)$). Indeed, the efficiency of the ABC-MCMC (ABC-IS) algorithm relies on $q(\theta)$ ($f(\theta)$), and a poor choice may lead to misleading results. Intuition suggests that $q(\theta)$ ($f(\theta)$) should be as similar as possible to the posterior distribution. However, as the likelihood function is unavailable, so is the posterior distribution, and this makes the determination of $q(\theta)$ ($f(\theta)$) difficult. Notice that ABC algorithms based on Sequential Monte Carlo (SMC) methods, such as those proposed by Sisson *et al.* (2007, 2009), and by Beaumont *et al.* (2009) can be seen as a generalization of ABC-IS, where a perturbing kernel is used in order to bring the simulated values as close as possible to the target, by progressively reducing the threshold ϵ. The choice of the perturbing kernel plays a crucial role in the performance of these algorithms.

In this section we discuss a default proposal distribution, which can be used as an independent kernel for ABC-MCMC or like an importance density for ABC-IS. Given a summary statistic $\eta(\cdot)$, which is assumed to be informative, although not necessarily sufficient for θ, the relation between η and θ is considered. In the scalar parameter case, by treating this relation as an unbiased estimating function and using the theory of quasi-likelihoods (McCullagh, 1991), we derive a normal distribution – with suitable parameters depending on the observed summary statistics – on the space of η. Finally, using the aforementioned relation between η and θ, which is assumed to be one-to-one,

we transform the proposed values drawn from the normal distribution, to θ, leading to proposed parameter values θ^*. For vector-valued parameters, quasi-likelihoods are difficult to find, and in such cases we present an extension based on asymptotic arguments.

This proposal, which we call the QL proposal, can be used as an independent kernel for ABC-MCMC or as importance density for ABC-IS. The QL proposal has the obvious advantage, over other type of proposals, in that it is built upon the relation between η and θ and takes the observed data into account. Hence, QL produces candidate values which are in the bulk of the likelihood. However, QL is effective if $\eta(\cdot)$ is informative for θ, as is the case for every ABC algorithm. Moreover, it is assumed there exists a one-to-one relation between them. In practice, this function may be unknown, and in the following we show how it can be to estimated by means of usual regression techniques.

4.2.1 The quasi-likelihood proposal

The theory and the use of estimating equations and of the related quasi- and quasi-profile likelihood functions has received much attention in recent years; see among others, Liang & Zeger (1995); Barndorff-Nielsen (1995); Desmond (1997); Heyde (1997); Adimari & Ventura (2002); Severini (2002); Wang & Hanfelt (2003); Bellio *et al.* (2008). See, in addition, Ventura *et al.* (2010); Lin (2006); Greco *et al.* (2008) discuss the use of quasi-likelihood functions in the Bayesian setting.

Let $\eta \in \mathbb{R}$, let y be a realization of $Y \sim p(y;\theta)$ with $\theta \in \mathbb{R}^d$. Moreover, let $s(\theta;\eta)$ be an unbiased estimating function, based on the data η, *i.e.* $E_\theta\{s(\theta;\eta)\} = 0$.

The quasi-likelihood for θ, based on $s(\theta;\eta)$ is given by (McCullagh, 1991)

$$L_Q(\theta) \;=\; \exp\left\{\int_{c_0}^{\theta} A(x)s(x;\eta)\,\mathrm{d}x\right\}, \tag{4.5}$$

where $A(\theta) = K_s(\theta)/J_s(\theta)$, $J_s(\theta) = \mathrm{var}_\theta\{s(\theta;\eta)\}$, $K_s(\theta) = E_\theta\{-ds(\theta;\eta)/d\theta\}$, and c_0 is an arbitrary constant.

For a scalar parameter (4.5) is easy to compute. The aim of this section is to use (4.5) to derive a proposal distribution, as shown by the following proposition.

Proposition 4.1. *Let $\psi(\theta) = E_\theta(\eta;\theta)$ be a bounded regression function under the full model $p(y;\theta)$, for which $|\psi'(\theta)| < \infty$, where $\psi'(\theta) = d\psi(\theta)/d\theta$. Moreover, let the variance $\sigma^2_\psi = var_\theta(\eta;\theta)$ be constant with respect to θ.*

Consider the estimating function $s(\theta;\eta) = \eta - \psi(\theta)$. Then

$$L_Q(\theta) = \phi\left\{\frac{\psi(\theta) - \eta}{\sigma_\psi}\right\},$$

where $\phi(\cdot)$ is the the standard normal density function.

Proof. The estimating function $s(\theta;\eta)$ is unbiased since by definition $E\{\eta - \psi(\theta)\} = \psi(\theta) - \psi(\theta) = 0$. In this case $J_s(\theta) = \psi'(\theta)$, $K_s(\theta) = \sigma^2_\psi$, so $A(\theta) = \psi'(\theta)/\sigma^2_\psi$ and from (4.5) we have that

$$\begin{aligned} L_Q(\theta) &= \exp\left[\int_{c_0}^{\theta} \frac{\psi'(\theta)}{\sigma^2_\psi}\{\eta - \psi(\theta)\}\right] \\ &\propto \frac{1}{\sigma_\psi}\exp\left[-\frac{\{\psi(\theta)-\eta\}^2}{2\sigma^2_\psi}\right], \end{aligned} \tag{4.6}$$

which is the kernel of the normal distribution centred at η with variance σ^2_ψ. □

Following Proposition 4.1, we suggest to use the quasi-likelihood (4.6), with $\eta = \eta^{\text{yobs}}$, as a proposal distribution for θ, which is given by

$$q(\theta) = L_Q(\theta)|\psi'(\theta)| = \phi\left\{\frac{\psi(\theta) - \eta^{\text{obs}}}{\sigma_\psi}\right\}|\psi'(\theta)|. \tag{4.7}$$

The distribution (4.7) can be used as an independent kernel within the ABC-MCMC algorithm, as shown for instance by Algorithm 8. Similarly, (4.7) can by used as an importance density in Algorithm 7. However, in the following we focus on the use of (4.7) as a proposal distributions for ABC-MCMC algorithms.

Result: A dependent sample $(\theta^{(1)}, \ldots, \theta^{(m)})$ from $\pi(\theta|\eta^{\text{obs}})$
Data: ϵ, $\eta(\cdot)$
1 set $\theta^{(0)} = \psi^{-1}(\eta^{\text{obs}})$
for $t = 1 \to m$ **do**
  repeat
2 draw $\psi^* \sim N(\eta^{\text{obs}}, \sigma^2_\psi)$
3 set $\theta^* = \{\theta : \psi^{-1}(\psi^*) = \theta^*\}$
4 draw $y \sim p(y;\theta^*)$ and set $\eta = \eta(y)$
5 draw $u \sim U(0,1)$
6 compute $R = \frac{\pi(\theta^*)L_Q(\theta^{(t-1)})|\psi'(\theta^{(t-1)})|}{\pi(\theta^{(t-1)})L_Q(\theta^*)|\psi'(\theta^*)|}$
  until $\rho(\eta, \eta^{obs}) \leq \epsilon$ *and* $u \leq R$;
7 set $\theta^{(t)} = \theta^*$
end

Algorithm 8: ABC-MCMC with the quasi-likelihood proposal.

Essentially, the regression function $\psi(\theta)$ acts as a suitable reparametrization of θ, which requires s and θ to be stochastically related. Except in some situations (see for instance

Sect. 4.1), $\psi(\theta)$ as well as σ_ψ are generally unknown. Hence, we suggest to replace them by their estimated versions $\hat{\psi}(\theta)$ and $\hat{\sigma}_\psi$, obtained as follows.

Estimation of $\psi(\theta)$

The estimation of $\psi(\theta)$ and σ_ψ can be performed in a pilot-run simulation. In this pilot-run study, we set an equispaced grid of M values of θ denoted with $\theta_p = (\theta_p^{(1)}, \ldots, \theta_p^{(M)})$, suitably taken in some large subset of Θ. We simulate a dataset for each parameter value form the full model $p(y;\theta)$, and end up with an M-dimensional vector of simulated summary statistics $\eta_p = (\eta_p^{(1)}, \ldots, \eta_p^{(M)})$. Next, η_p is regressed on θ_p, and set the estimated regression function equal to $\hat{\psi}(\theta)$ and σ_ψ^2 is estimated by the residual variance $\hat{\sigma}_\psi^2 = M^{-1}\sum_{i=1}^{M}\{\hat{\psi}(\theta_p^{(i)}) - \eta_p^{(i)}\}$.

The regression estimator $\hat{\psi}(\theta)$ can be any method which provides smoothed functions, which are at least once differentiable. In the example shown in the next section, we consider smoothing splines, for which the required first derivative, *e.g.* the Jacobian of the transformation, can be readily obtained. The inverse $\hat{\psi}^{-1}(\psi^*)$, at point ψ^*, can be computed by usual numerical methods, such as the bisection method.

The range of the grid θ_p must be wide enough to include values of η_p which are compatible with the observed one η^{obs}. Lastly, the hypothesis of constant regression variance $\sigma_\psi^2 = \text{var}_\theta(\eta;\theta)$, can be inspected by usual regression diagnostics, and if not satisfied can be achieved by suitable transformations.

Generalization for $d > 1$

Now let $d > 1$, and let η^{obs} be the d-dimensional vector of suitable summary statistics. In this case the theory of quasi-likelihoods is not very helpful. Indeed, when $d > 1$ it is known that $L_Q(\theta)$ exists if and only if the matrix $B(\theta)$ is symmetric (see, *e.g.*, McCullagh, 1991).

We propose to take the quadratic form

$$L_{mQ}(\theta) = |\Sigma_\psi|^{-1/2} \exp\left[-\frac{1}{2}\{\psi(\theta) - \eta^{\text{obs}}\}^T \Sigma_\psi^{-1}\{\psi(\theta) - \eta^{\text{obs}}\}\right] \tag{4.8}$$

as a quasi-likelihood for vector-valued parameters. In (4.8), $\psi(\theta) = (\psi_1(\theta), \ldots, \psi_d(\theta))$ is a bounded and monotone vector-valued regression function and Σ_ψ is the conditional covariance matrix assumed to be independent of θ.

Therefore, the proposal distribution for ABC-MCMC is

$$q(\theta) = L_{mQ}(\theta)|\psi'(\theta)| = \phi_d(\psi(\theta); \eta^{\text{obs}}, \Sigma_\psi)|\psi'(\theta)|. \tag{4.9}$$

The functions $\psi(\theta)$ and the matrix Σ_r are generally unknown. To estimate them, we suggest to consider a pilot-run simulation study as before. In particular, let θ_{pd} be an $M^d \times d$ matrix given by the Cartesian product of d regular grids $\theta_{p1}, \dots \theta_{pd}$, each made of M equispaced values, in some suitable subspace of Θ. For each value of θ_{pd}, we take the associated η simulated from the model, and consider the matrix of simulated values η_{pd}. The next step is to regress η_{pi} on θ and take as $r_i(\theta)$ the estimated regression function $\hat{r}_i(\theta)$ $(i = 1, \dots, d)$. Moreover, given $e = (e_1, \dots, e_d)$, the $M^d \times d$ matrix of regression residuals, we approximate Σ_ψ by $\hat{\Sigma}_\psi = M^{-1} e^T e$.

Many of the observations made previously apply also here, suitably adapted. In order to guarantee enough flexibility, and since we are mainly interested in predicting η, $r(\theta)$ can be considered in the class of generalized additive regression models (Stone, 1985), in which each of the d components of θ enters the linear predictor by means of a smoothing spline as discussed, for instance, in Faraway (2006, Ch. 12). While the proposed approach limits the number of statistics to be equal to the number of parameters, this is in line with the general recommendation in the ABC literature.

4.2.2 An example: the coalescent model

Given a set of n DNA sequences, the aim of the coalescent model (Tavaré *et al.*, 1997) is to estimate the effective mutation rate $\theta > 0$, under the infinitely-many-sites assumption. In this model the mutations occur at rate θ at DNA sites that have not been hit by mutation before. If a site is affected by a mutation, then it is said to be segregating in the sample. In this example, the summary statistic is the number of segregating sites, *e.g.* $\eta = y$ (see also Blum & François, 2010) The generating mechanism for y is the following:

(1) generate the length of the genealogical tree of n sequences, given by $T_n = \sum_{j=2}^{n} W_j$, where W_j are exponential random variables with mean $2/j(j-1)$;

(2) generate $Y \sim \text{Poi}(\theta T_n/2)$, from the Poisson distribution with mean $\theta T_n/2$.

The likelihood of the coalescent model is the marginal density of $Y|\theta$ with T_n integrated out. This likelihood has a closed form only for $n = 2$. However, an approximation of the likelihood for every n, can be obtained by simply integrating out T_n via standard Monte Carlo integration.

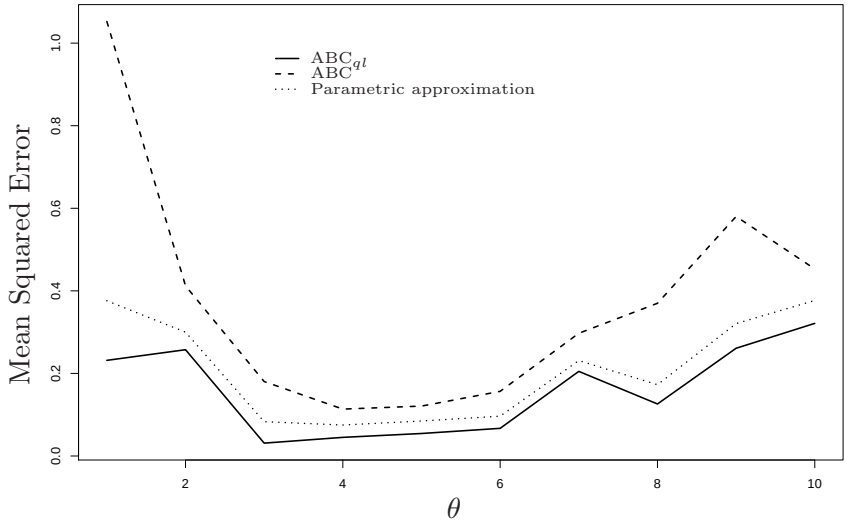


FIGURE 4.9: Coalescent model. Comparisons of MSEs calculated for different values of θ over 100 replications, for the mean of the parametric posterior, the ABC and ABC$_{ql}$.

In this example, the prior is assumed to be an exponential random variable with unit mean, the parameter is taken in logarithmic scale, and the summary statistic is $\eta = \log(y + 1)$. The function $\psi(\theta)$ is estimated with a smoothing spline in a grid of $m = 1000$ values, and the related Jacobian is computed numerically. The required inversions are performed with the bisection method.

In Figure 4.9 we compare the ABC-MCMC method with the proposed kernel based on the quasi-likelihood, called ABC$_{ql}$, with the standard ABC accept/reject algorithm, as well as the parametric posterior, where the likelihood is obtained via Monte Carlo integration. As an example, we take a sample of $n = 100$ sequences and consider the MSE over 100 replications from the model and with different parameter values. At each replication, the ABC abd ABC$_{ql}$ posteriors are approximated by a simulated sample of size 1000 obtained by setting ϵ to 0.1%th quantile of the absolute distances of the summary statistics.

By treating the parametric approximation as the gold standard, we compare the quantiles of ABC and ABC$_{ql}$, respectively with those of the parametric approximation Q_p^0, by the relative difference $(Q_p - Q_p^0)/Q_p^0$, where Q_p is the pth quantile of ABC or ABC$_{ql}$ (see also Blum & François, 2010), for $p \in (0, 1)$. Figure 4.10 shows such relative differences, where we can see that these differences are more robust with respect to θ for ABC$_{ql}$ rather than for the ABC, which can be explained by the impact of the prior in the standard ABC algorithm.

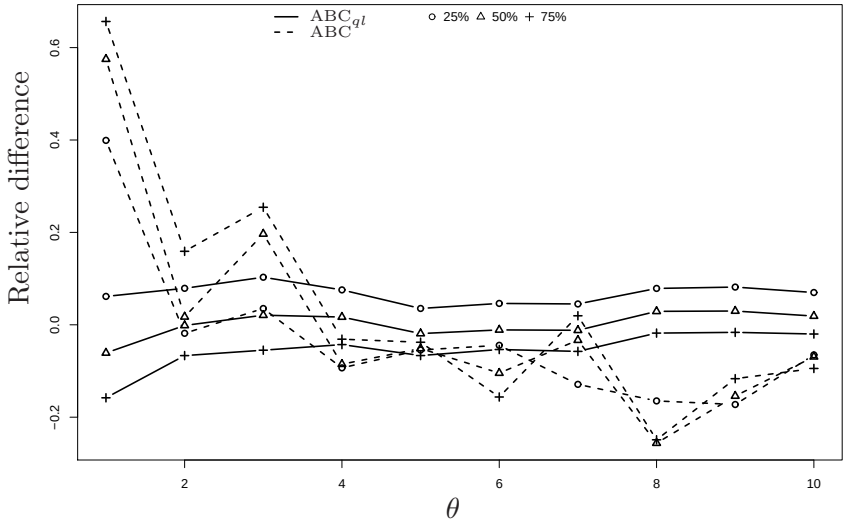


FIGURE 4.10: Coalescent model. Comparison of ABC and ABC$_{ql}$ against the parametric approximation in terms of relative differences between the quantiles.

4.2.3 Remarks

We discussed a new proposal distribution for ABC-MCMC algorithms. Given a summary statistic and provided it is informative for θ, the proposal distribution in the scalar parameter case is obtained by using the theory of quasi-likelihood. The idea of the quasi-likelihood proposal is readily extended for dealing with multidimensional parameters and its application to real examples is under development.

A crucial point to the success of the method is the existence of a stochastic relation between η and θ, which we assume to be invertible. This relation is typically unknown, and we proposed an estimated version obtained by means of smoothing splines.

As a final comment, we remark that the proposed kernel can be used also as an importance density for ABC-IS algorithms, although we did not pursue this use here.

Bibliography

Adimari, G. & Ventura, L. (2002). Quasi-profile log likelihoods for unbiased estimating functions. *Annals of the Institute of Statistical Mathematics* **54**, 235–244.

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

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Barndorff-Nielsen, O. (1995). Quasi profile and directed likelihoods from estimating functions. *Annals of the Institute of Statistical Mathematics* **47**, 461–464.

Barndorff-Nielsen, O. & Chamberlin, S. (1994). Stable and invariant adjusted directed likelihoods. *Biometrika* **81**, 485–499.

Barndorff-Nielsen, O. E. & Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*. London: Chapman and Hall.

Barndorff-Nielsen, O. E. & Cox, D. R. (1994). *Inference and Asymptotics*. Boca Ranton, Florida: Chapman & Hall/CRC.

Barthelmé, S. & Chopin, N. (2011). Expectation-propagation for likelihood-free inference. *arXiv preprint* **arXiv:1107.5959**.

Bates, D. M. & Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley Online Library.

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.

Bellio, R., Greco, L. & Ventura, L. (2008). Modified quasi-profile likelihoods from estimating functions. *Journal of Statistical Planning and Inference* **138**, 3059–3068.

Bernardo, J. M. & Smith, A. F. M. (1994). *Bayesian Theory*. New York: Wiley.

Besag, J. L. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B* **36**, 192–236.

Blum, M. & François, O. (2010). Non-linear regression models for Approximate Bayesian Computation. *Statistics and Computing* **20**, 63–73.

Brazzale, A. R. & Davison, A. C. (2008). Accurate parametric inference for small samples. *Statistical Science* **23**, 465–484.

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

Chang, H., Kim, B. & Mukerjee, R. (2009). Bayesian and frequentist confidence intervals via adjusted likelihoods under prior specification on the interest parameter. *Statistics: A Journal of Theoretical and Applied Statistics* **43**, 203–211.

Chang, H. & Mukerjee, R. (2006). Probability matching property of adjusted likelihoods. *Statistics and Probability Letters* **76**, 838–842.

Chen, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *Journal of the American Statistical Association* **89**, 818–824.

Chen, M.-H., Shao, Q.-M. & Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321.

Chib, S. & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association* **96**, 270–281.

Chipman, H., George, E. I. & McCulloch, R. E. (2001). The practical implementation of Bayesian model selection. In *Model Selection*, P. Lahiri, ed. Lecture Notes–Monograph Series vol. 38. Beachwood, OH: Institute of Mathematical Statistics, pp. 65–116.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society: Series B* **34**, 187–220.

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62**, 269–276.

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

Cox, D. R. & Wermuth, N. (1990). An approximation to maximum likelihood estimates in reduced models. *Biometrika* **77**, 747–761.

Datta, G. S. & Mukerjee, R. (2004). *Probability Matching Priors: Higher Order Asymptotics*. Lecture Notes in Statistics vol. 178. New York: Springer-Verlag.

Davison, A. C. (2003). *Statistical Models*. Cambridge: Cambridge University Press.

Davison, A. C., Fraser, D. A. & Reid, N. (2006). Improved likelihood inference for discrete data. *Journal of the Royal Statistical Society: Series B* **68**, 495–508.

DeBruijn, N. G. (1961). *Asymptotic Methods in Analysis*. New York: Dover Publications.

Desmond, A. (1997). Optimal estimating functions, quasi-likelihood and statistical modelling. *Journal of Statistical Planning and Inference* **60**, 77–104.

DiCiccio, T. J., Field, C. A. & Fraser, D. (1990). Approximations of marginal tail probabilities and inference for scalar parameters. *Biometrika* **77**, 77–95.

DiCiccio, T. J., Kass, R. E., Raftery, A. & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association* **92**, 903–915.

DiCiccio, T. J. & Martin, M. A. (1991). Approximations of marginal tail probabilities for a class of smooth functions with applications to bayesian and conditional inference. *Biometrika* **78**, 891–902.

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.

Evans, M. & Swartz, T. (1995). Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science* **10**, 254–272.

Evans, M. & Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press.

Faraway, J. J. (2006). *Extending the Linear Model with R*. New York: Springer.

Fassina, A., Cappellesso, R., Guzzardo, V., Dalla Via, L., Piccolo, S., Ventura, L. & Fassan, M. (2011). Epithelial–mesenchymal transition in malignant mesothelioma. *Modern Pathology* **25**, 86–99.

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.

Fraser, D. A. S., Reid, N. & Wu, J. (1999). A simple general formula for tail probabilities for frequentist and Bayesian inference. *Biometrika* **86**, 249–264.

Friel, N. & Wyse, J. (2012). Estimating the evidence – a review. *Statistica Neerlandica* **66**, 288–308.

Gelfand, A. E. & Dey, D. K. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B* **56**, 501–514.

Gelfand, A. E. & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–405.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2003). *Bayesian Data Analysis*. Boca Ranton, Florida: Chapman & Hall/CRC.

Genz, A. & Bornkamp, B. (2011). *bayespack: Numerical Integration for Bayesian Inference*. R package version 1.0-2.

Gilbert, P. & Varadhan, R. (2012). *numDeriv: Accurate Numerical Derivatives*. R package version 2012.9-1.

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

Greco, L., Racugno, W. & Ventura, L. (2008). Bayesian analysis in regression models using pseudo-likelihoods. *Journal of Statistical Inference and Planning* **138**, 1258–1270.

Guihenneuc-Jouyaux, C. & Rousseau, J. (2005). Laplace expansions in Markov chain Monte Carlo algorithms. *Journal of Computational and Graphical Statistics* **14**, 75–94.

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.

Heyde, C. (1997). *Quasi-Likelihood and its Application*. New York: Springer Verlag.

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

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

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001). *Bayesian Survival Analysis*. New York: Springer.

Jones, M. (2002). Marginal replacement in multivariate densities, with application to skewing spherically symmetric distributions. *Journal of Multivariate Analysis* **81**, 85–99.

Kalbfleisch, J. D. (1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society: Series B* **40**, 214–221.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kass, R. E., Tierney, L. & Kadane, J. B. (1989). Approximate methods for assessing influence and sensitivity in Bayesian analysis. *Biometrika* **76**, 663–674.

Kass, R. E., Tierney, L. & Kadane, J. B. (1990). The validity of posterior expansions based on laplace's method. In *Bayesian and Likelihood Methods in Statistics and Econometrics: Essays in Honor of George A. Barnard*, S. Geisser, J. Hodges, S. Press & A. Zellner, eds. North Holland.

Kass, R. E. & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association* **91**, 1343–1370.

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

Kharroubi, S. A. & Sweeting, T. J. (2010). Posterior simulation via the signed root log-likelihood ratio. *Bayesian Analysis* **5**, 787–815.

Kim, Y. & Kim, D. (2009). Bayesian partial likelihood approach for tied observations. *Journal of Statistical Planning and Inference* **139**, 469–477.

Lahiri, P. (2001). *Model Selection*. Lecture Notes–Monograph Series vol. 38. Beachwood, OH: Institute of Mathematical Statistics.

Lavine, M. & Schervish, M. J. (1999). Bayes factors: what they are and what they are not. *The American Statistician* **53**, 119–122.

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

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.

Lewis, S. M. & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association* **92**, 648–655.

Liang, K. & Zeger, S. (1995). Inference based on estimating functions in the presence of nuisance parameters. *Statistical Science* **10**, 158–173.

Lin, L. (2006). Quasi Bayesian likelihood. *Statistial Methodology* **3**, 444–455.

Lindley, D. V. (1961). The use of prior probability distributions in statistical inference and decision. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. 453–468.

Lindley, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint. Part 2: Inference*. Cambridge: Cambridge University Press.

Lindley, D. V. (1980). Approximate Bayesian methods. *Trabajos de Estadistica Y de Investigacion Operativa* **31**, 223–245.

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.

Marin, J.-M. & Robert, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York: Springer.

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.

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

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.

Naylor, J. D. & Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society: Series C* **31**, 214–225.

Newton, M. A. & Raftery, A. E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society: Series B* **56**, 3–48.

Noble, B. & Daniel, J. W. (1988). *Applied Linear Algebra*. New York: Prentice-Hall.

Nott, D. J., Fielding, M. & Leonte, D. (2009). On a generalization of the Laplace approximation. *Statistics & Probability Letters* **79**, 1397–1403.

O'Hagan, A. & Forster, J. J. (2004). *Kendall's Advanced Theory of Statistics. Volume 2B: Bayesian Inference*. London: Arnold, 2nd ed.

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

Pace, L. & Salvan, A. (2006). Adjustments of the profile likelihood from a new perspective. *Journal of Statistical Planning and Inference* **136**, 3554–3564.

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.

Perrakis, K., Ntzoufras, I. & Tsionas, E. G. (2013). On the use of marginal posteriors in marginal likelihood estimation via importance-sampling. *ArXiv e-prints* **1311.0674**.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Racugno, W., Salvan, A. & Ventura, L. (2010). Bayesian analysis in regression models using pseudo-likelihoods. *Communications in Statistics - Theory and Methods* **39**, 3444–3455.

Raftery, A. E., Madigan, D. & Volinsky, C. T. (1996). Accounting for model uncertainty in survival analysis improves predictive performance. In *Bayesian Statistics*, J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds., vol. 5. Oxford Science Publications, pp. 323–349.

Reid, N. (1996). Likelihood and Bayesian approximation methods. In *Bayesian Statistics*, J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds., vol. 5. Oxford Science Publications, pp. 351–368.

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

Reid, N. & Sun, Y. (2010). Assessing sensitivity to priors using higher order approximations. *Communications in Statistics: Theory and Methods* **39**, 1373–1386.

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

Robert, C. P. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. New York: Springer.

Robert, C. P. & Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer, 2nd ed.

Rubio, F. J. & Johansen, A. M. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics* **7**, 1632–1654.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B* **71**, 319–392.

Severini, T. (2002). Modified estimating functions. *Biometrika* **89**, 333–343.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. New York: Oxford University Press.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Sinha, D., Ibrahim, J. G. & Chen, M.-H. (2003). A Bayesian justification of Cox's partial likelihood. *Biometrika* **90**, 629–641.

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.

Skovgaard, I. M. (2001). Likelihood asymptotics. *Scandinavian Journal of Statistics* **28**, 3–32.

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.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics* **13**, 435–844.

Sweeting, T. (1996). Approximate Bayesian computation based on signed roots of log-density ratios (with discussion). In *Bayesian Statistics*, J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith, eds., vol. 5. Oxford Science Publications, pp. 427–444.

Sweeting, T. J. (1995). A framework for Bayesian and likelihood approximations in statistics. *Biometrika* **82**, 1–23.

Sweeting, T. J. (1999). On the construction of Bayes–confidence regions. *Journal of the Royal Statistical Society: Series B* **61**, 849–861.

Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. J. (1997). Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–18.

Tibshirani, R. J. (1989). Noninformative priors for one parameter of many. *Biometrika* **76**, 604–608.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

Tierney, L., Kass, R. E. & Kadane, J. B. (1989). Approximate marginal densities of nonlinear functions. *Biometrika* **76**, 425–433.

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.

Ventura, L., Cabras, S. & Racugno, W. (2009). Prior distributions from pseudo-likelihoods in the presence of nuisance parameters. *Journal of the American Statistical Association* **104**, 768–774.

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

Ventura, L., Sartori, N. & Racugno, W. (2013). Objective Bayesian higher-order asymptotics in models with nuisance parameters. *Computational Statistics & Data Analysis* **60**, 90–96.

Volinsky, C. T., Madigan, D., Raftery, A. E. & Kronmal, R. A. (1997). Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. *Journal of the Royal Statistical Society: Series C* **46**, 433–448.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society: Series B* **31**, 80–88.

Wang, M. & Hanfelt, J. (2003). Adjusted profile estimating function. *Biometrika* **90**, 845–858.

Wuertz, D., many others & see the SOURCE file (2013). *fCopulae: Rmetrics - Dependence Structures with Copulas*. R package version 3000.79.

Yin, G. (2009). Bayesian generalized method of moments. *Bayesian Analysis* **4**, 191–207.

Erlis Ruli

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 049 827 4111
Mobile: +39 329 8149214
e-mail: ruli@stat.unipd.it

Current Position

Since January 2011; (expected completion: December 2013)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Recent Advances in Approximate Bayesian Computation Methods
Supervisor: Prof. Laura Ventura
Co-supervisor: Prof. Nicola Sartori.

Research interests

- Monte Carlo methods in statistics such as Markov chain Monte Carlo, Importance sampling, Sequential Monte Carlo methods from a methodological and applied perspective.
- ABC, pseudo-likelihood, estimating functions and scoring rules
- Asymptotic computations
- Model choice in regular and ultra high dimensions
- Mixture models, Hidden Markov models, Survival analysis, Spatio-temporal data analysis, Causal inference, Big data.

Education

October 2007 – July 2010
Master degree (*laurea specialistica/magistrale***) in Economic Sciences**.
University of Cagliari, Faculty of Economics
Title of dissertation: "Why so many/few new firms? A Bayesian analysis with random effects" (in Italian)
Supervisor: Prof. Walter Racugno
Final mark: 110/110 cum laude

October 2003 – November 2006
Bachelor degree (*laurea triennale***) in European Economics and Politics**.
University of Cagliari, Faculty of Political Sciences
Title of dissertation: "The covered interest rate parity: econometric analysis and empirical evidence" (in Italian)
Supervisor: Prof. Emaluela Marrocu
Final mark: 110/110.

Visiting periods

May 2012
Universidad Carlos III de Madrid,
Getafe, Spain.
Supervisor: Prof. Stefano Cabras

November 2012 – December 2012
Centre de Richerche en Economie et Statistique,
Paris, France .
Supervisor: Prof. Christian P. Robert

Work experience

June 2007 – September 2007
Osservatorio Economico della Sardegna.
Worked in data analysis and data processing for the "Industial Synthetic Reports" project (Schede sintetiche di settore).

Computer skills

- Advanced R user
- Basic C/C++ user
- Basic user of Mathematica, Matlab, Stata and EViews.

Language skills

Albanian: native; Italian: fluent (written and spoken); English: fluent (written and spoken);

Publications

Ruli, E., Sartori, N., Ventura, L. (2013). A note on marginal posterior simulation via higher-order tail area approximations. *Bayesian Analysis*, doi: 10.1212/13-BA851.

Ventura, L., Ruli, E., Racugno, W. (2013). A note on approximate Bayesian credible sets based on modified loglikelihood ratios. *Statistics and Probability Letters* **83**, 2467–2472.

Working papers
Ruli, E., Ventura, L., (2013). Higher-order Bayesian approximations for pseudo-posterior distributions. submitted.

Cabras, S., Castellanos, M.E., Ruli, E., (2013). Approximate Bayesian computation with quasi-likelihoods. under revision.

Ruli, E., Sartori, N., Ventura, L., (2013). Approximate Bayesian Computation with composite score functions. submitted. submitted.

Cabras, S., Castellanos, M. E., Ruli, E., (2013). A Quasi likelihood approximation of posterior

distributions for likelihood-intractable complex models. submitted.

Ruli, E., Ventura, L., Racugno, W., (2013). A note on default Bayesian inference for the consensus mean in inter-laboratory studies. submitted.

Conference presentations

Ruli, E., Ventura, L., (2013). Advances in approximate Bayesian computation with modified pseudo-likelihood roots. (poster) *S.Co.2013*, Milano, Italy, September 9-11, USB stick (ISBN:9788864930190).

Ventural, L., Ruli, E., Racugno, W., (2013). Default Bayesian inference for the consensus mean in inter-laboratory studies. (contributed) *S.Co.2013*, Milano, Italy, September 9-11, USB stick (ISBN:9788864930190), 1-6.

Ruli, E., Ventura, L., Racugno, W., (2013). Approximate Bayesian inference based on modified log-likelihood ratios. (contributed) *28th IWSM*, Palermo, Italy, July, 8-12.

Cabras, S., Castellanos, M.E., Ruli, E., (2012), ABC-MCMC algorithms with quasi-likelihoods. (contributed) *CFE-ERCIM 2012*, Oviedo, Spain, Dicember 1-3 (ISBN: 978-84-937882-2-1).

Ruli, E., Ventura, L. (2012), Bayesian approximation methods for pseudo-posterior distributions in the presence of nuisance parameters. (poster) *27th IWSM*, Prague, Czech Republic, July, 16-20.

Ruli, E., Ventura, L. (2013). Bayesian marginal posterior simulation from the signed likelihood root: an applications to the Cox regression model. (contributed) *Proceedings of the IX National Meeting of the Italian Biometric Society*, Bressanone, Italy, June 27-28.

Ruli, E., Ventura, L. (2012), Modern Bayesian inference in zero-inflated Poisson models. (contributed) *Atti della XLVI Riunione Scientifica della SIS*, Roma, Italy, June 20-22, USB stick (ISBN: 978-88-6129-882-8).

Teaching experience

March 2010 – December 2010
Social Statistics, Economics, Political Economics, Macroeconomics
Bachelor
Teaching task (exercises and lab), total number of hours 80
Faculty of Political Sciences, University of Cagliari

March 2010 – July 2010
Statistics
Bachelor
Teaching task (exercises and lab), total number of hours 30
Faculty of Economics, Universtiy of Cagliari
Instructor: Prof. Walter Racugno

May 2009 – December 2009
Social Statistics, Economics, Political Economics, Macroeconomics
Bachelor
Teaching task (exercises and lab), total number of 80

Faculty of Political Sciences, Universtiy of Cagliari

March 2009 – July 2009
Statistics
Teaching task (exercises and lab), total number of hours 30
Faculty of Economics, Universtiy of Cagliari
Instructor: Dr. Stefano Cabras

November 2008 – February 2009
Economics
Teaching task (exercises), total number of hours 30
Faculty of Political Sciences, Universtiy of Cagliari
Instructor: Prof. Sergio Lodde

References

Prof. Walter Racugno
Institution: Department of Mathematics and Informatics, University of Cagliari
Address: Via Ospedale, 72, 09124 Cagliari, Italy
Phone: +39 070675-8532
e-mail: racugno@unica.it

Dr. Stefano Cabras
Institution: Department of Mathematics and Informatics, University of Cagliari
Address: Via Ospedale, 72, 09124 Cagliari, Italy
Phone: +39 070675-8535
e-mail: s.cabras@unica.it

Prof. Laura Ventura
Institution: Department of Statistical Sciences, University of Padova
Address: Via Cesare Battisti, 241, 35121 Padova, Italy
Phone: +39 049 8274177
e-mail: ventura@stat.unipd.it

Prof. Nicola Sartori
Institution: Department of Statistical Sciences, University of Padova
Address: Via Cesare Battisti, 241, 35121 Padova, Italy
Phone: +39 049 8274127
e-mail: sartori@stat.unipd.it

Padova,
28 January 2014

Erlis Ruli