Bayesian Monte Carlo Testing With One-dimensional Measures of Evidence

Ivair R. Silva · **Reinaldo Marques**

Abstract Bayesian hypothesis testing procedures are constructed by means of test statistics which are functions of the posterior distribution. Usually, the whole sample vector is selected to form the sufficient empirical part of the posterior distribution. But, in certain problems, one may prefer to use well-established one-dimensional sufficient statistics in place of the sample vector. This paper introduces a Bayesian Monte Carlo procedure specially designed for such cases. It is shown that the performance of this new approach is arbitrarily close to the exact Bayesian test. In addition, for arbitrary desired precisions, we develop a theoretical rule of thumb for choosing the minimum number m_0 of Monte Carlo simulations. Surprisingly, m_0 does not depend on the shape of loss/cost functions when those are used to compound the test statistic. The method is illustrated for testing mean vectors in high-dimension and for detecting spatial clusters of diseases in aggregated maps.

Keywords Bayes Factor · Expected loss · Exact test · Measure of evidence · Monte Carlo

1 Introduction

A statistical hypothesis is any statement about an unknown but fixed population parameter $\theta \in \mathbb{R}^N$, where N is a positive integer. In hypothesis testing, the goal is to decide which of two hypotheses is true. The usual format of the hypotheses is $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 form a partition of Θ, the parameter space.

In general, hypothesis testing demands to collect a n-dimensional sample $\tilde{\mathbf{X}} = (X_1, \cdots, X_n)$, which, supposedly, contains information about θ. But, before having a realization $\tilde{\mathbf{x}}_0$ of $\tilde{\mathbf{X}}$, the analyst might have some idea of what are the most and the less plausible values for θ. Such a subjective uncertainty about θ can be mathematically translated in terms of a probability measure, the so called 'prior distribution'.

I. Silva
Department of Population Medicine, Harvard Medical School and Harvard
Pilgrim Health Care Institute, Boston, MA, USA
Department of Statistics, Federal University of Ouro Preto, MG, Brazil
E-mail: ivairest@gmail.com

R. Marques
Department of Statistics, University of Oslo, Oslo, Norway
Actuarial Division, Federal University of Alfenas, Brazil

In contrast to the widely used notation, '$\pi(\theta)$', here the prior distribution is instead denoted by $\pi_\theta(y)$, $y \in \mathbb{R}^N$. This new notation is more consistent with the fact that the argument, y, of the prior distribution, and the parameter to be inferred, θ, are by definition two different apparatus. This is useful to avoid the misleading interpretation that θ is random in Bayesian analysis. Such clarification is important for a proper comprehension of the reasoning developed in this paper.

The exact Bayesian test consists on using a sufficient statistic for θ, say $T(\tilde{\mathbf{X}})$, in order to update the analyst's uncertainty about the plausibility of each hypothesis. By 'sufficient statistic' we mean that $T(\tilde{\mathbf{X}})$ can be any mapping of $\tilde{\mathbf{X}}$ (not involving θ) from $\mathbb{R}^n$ to $\mathbb{R}^m$, with $m \leq n$, such that the conditional distribution of $\tilde{\mathbf{X}}$, given $T(\tilde{\mathbf{X}})$, does not depend of θ. For example, the usual sufficient statistic choice in Bayesian analysis is the whole sample, that is $T(\tilde{\mathbf{X}}) = \tilde{\mathbf{X}}$. But, the data reduction principle can be invoked to justify usage of real-valued sufficient statistics too, which is valid for Bayesian analysis without any logical or philosophical flaws.

Given an observed sample $\tilde{\mathbf{X}} = \tilde{\mathbf{x}}_0$, the updated uncertainty, called posterior distribution, is calculated through the Bayes rule as follows:

$$\pi_\theta(y|T(\tilde{\mathbf{x}}_0)) = f_T(T(\tilde{\mathbf{x}}_0)|y)\pi_\theta(y)/m(T(\tilde{\mathbf{x}}_0)), \tag{1}$$

where $f_T(T(\tilde{\mathbf{x}}_0)|y)$ is the likelihood with respect to $T(\tilde{\mathbf{x}}_0)$, but evaluated for $\theta := y$. The denominator is sometimes called 'predictive function', and it is given by:

$$m(T(\tilde{\mathbf{x}}_0)) = \int_{y \in \Theta} f_T(T(\tilde{\mathbf{x}}_0)|y)\pi_\theta(y)dy. \tag{2}$$

A Bayesian measure of evidence in favor of H_0, say $W(\tilde{\mathbf{X}})$, is then constructed as a function of $\pi_\theta(y|T(\tilde{\mathbf{x}}_0))$. Large values of $W(\tilde{\mathbf{X}})$ suggest that H_0 should be taken as true. Otherwise, H_1 should be taken as true. A well-known measure of evidence is the Bayes factor:

$$BF(T(\tilde{\mathbf{X}})|\pi_\theta) = \frac{Pr\left[\theta \in \Theta_0|T(\tilde{\mathbf{X}}),\pi_\theta\right]Pr\left[\theta \in \Theta_1|\pi_\theta\right]}{Pr\left[\theta \in \Theta_1|T(\tilde{\mathbf{X}}),\pi_\theta\right]Pr\left[\theta \in \Theta_0|\pi_\theta\right]}. \tag{3}$$

The probability $Pr\left[\theta \in \Theta_i|T(\tilde{\mathbf{X}}),\pi_\theta\right]$ in (3) refers to the posterior distribution given $T(\tilde{\mathbf{X}})$, while the probability $Pr\left[\theta \in \Theta_i|\pi_\theta\right]$ refers to the prior distribution $\pi_\theta(y)$, $i = 0, 1$. For precise hypotheses of the form $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, the prior distribution is conveniently defined as a discrete distribution, which leads to a discrete posterior distribution. Therefore, the Bayes factor can be calculated according to (3) for precise hypotheses too.

It is worth noting that, before observing the data, $BF(T(\tilde{\mathbf{X}})|\pi_\theta)$ is a random variable. Thus, $BF(T(\tilde{\mathbf{X}})|\pi_\theta)$ takes the place of a test statistic with decision set of the form:

$$\mathbf{R}(c) = \{b \in \mathbb{R} : b \leq c\}, \tag{4}$$

with $c > 0$ arbitrary. Therefore, H_1 is taken as true if $BF(T(\tilde{\mathbf{x}}_0)|\pi_\theta) \leq c$, otherwise, H_0 is taken as true. The choice of c can be guided by the scale suggested by Jeffrey (Jeffrey, 1961). Basically, if $-\log_{10}BF(T(\tilde{\mathbf{x}}_0)|\pi_\theta) > 2$ (same as $c \leq 0.01$), then H_1 is taken as true.

Alternatively, c can be established in a way to optimize some target performance measure, like, for example, to minimize the expected loss. The expected loss is a statistical pre-experimental performance measure used to evaluate the performance of test procedures. Section 2 offers a deeper discussion on this topic.

It merits remark that the posterior distribution is computable only if the probability distribution $F_T(t|\theta)$ of $T(\tilde{\mathbf{X}})$ has a known shape. If $F_T(t|\theta)$ is unknown, but samples of $T(\tilde{\mathbf{X}})$

can be somehow generated for fixed values of θ, Monte Carlo simulation can be used to perform the test. Monte Carlo plays a remarkable role in Bayesian analysis (Owen and Zhou, 2000; Besag, 2001; Pereira and Stern, 1999). The prominent methods designed specially to approximate the posterior distribution are based on *importance ratio* techniques, like for instance Importance Sampling, Metropolis-Hasting (MCMC), and particle methods (Chen et al, 2012; Robert and Casella, 2013). Likelihood-free techniques are another example of how important Monte Carlo is for Bayesian analysis, like the Approximate Bayesian Computation (ABC), (Marin et al, 2014; Fearnhead, 2017; Price et al, 2017), and the hybrid approaches ABC-MCMC, ABC-SMC, Resample-Move, Move-Reweighting, and Particle MCMC, (Gilks and Berzuini, 2001; Sisson et al, 2007; Andrieu et al, 2010; Didelot et al, 2011; Marques and Storvik, 2013; Everitt et al, 2017).

In this paper, we introduce a Bayesian Monte Carlo test that, unlike previous approaches, is not designed to approximate the exact test. The proposed method is a formal test procedure that works for any sample size, for arbitrary prior distributions, and it is quite simpler than Importance Sampling, Metropolis-Hasting, particle methods, and ABC algorithms. The main contribution of the present manuscript is to favor Bayesian hypothesis testing for problems where: (i) $T(\tilde{\mathbf{X}})$ is a real-valued function and (ii) the shape of the likelihood function with respect to $T(\tilde{\mathbf{X}})$ cannot be mathematically expressed, but Monte Carlo simulation for $T(\tilde{\mathbf{X}})$ is feasible.

As expected, the Monte Carlo variation affects the statistical performance of the Bayesian Monte Carlo test. To resolve this problem, a simple expression is derived in order to relate the number of simulations with the expected loss of the Monte Carlo approach relatively to the expected loss of the exact (although intractable) Bayesian test. With this, the performance of the proposed test can be made arbitrarily close to the exact Bayesian test using a pre-defined finite number of simulations. Hence, the main advantage of the proposed method is that it is not an approximation of the exact Bayesian test, neither requires asymptotic/convergence arguments to present good performance. The method is a formal hypothesis test procedure with performance arbitrarily close to the exact Bayesian test for any sample size.

The content of this paper is organized as follows: Next section offers an overview on expected loss and risk functions. Section 3 introduces the proposed Bayesian Monte Carlo test, and a road-map for applications involving tests of mean vectors in high dimension is offered. Section 4 contains a comparison study against three of the main competing methods found in the literature. Section 5 presents a real data example of application for spatial cluster detection in aggregated geographical maps. Section 6 closes the paper with a brief discussion.

2 Expected Loss

The statistical performance measure of interest in this paper is the so called 'expected loss', also known by 'risk function'. In decision theoretic analysis, the expected loss is a well-established performance measure for evaluation of hypothesis test procedures. The idea behind it is quite intuitive. The rationale is that some wrong decisions are more serious than others.

For example, consider hypotheses of the form $H_0 : \theta \geq \theta_0$ versus $H_1 : \theta < \theta_0$, where θ_0 is a fixed constant. If $\theta \geq \theta_0$ and H_1 is taken as true, a mistake has been made, i.e., the error of the Type I has occurred. If $\theta < \theta_0$ and H_0 is taken as true, a mistake has been made, i.e., the error of the Type II has occurred. The other possibilities represent the correct actions. If the Type I error occurs but θ is slightly bigger than θ_0, then the mistake may not be very

serious. But the mistake is possibly serious if θ is much bigger than θ_0. The intensity of how 'serious' a mistake is, for each possible value of θ, can be transmitted through a function of θ taking positive values in the real line, called 'loss function', and here denoted by $l_j(\theta)$, $j = 1,2$. For $j = 1$, $l_1(\theta)$ represents the loss associated to the Type I error event. For $j = 2$, $l_2(\theta)$ represents the loss associated to the Type II error event. The expected loss is a function of θ that combines the probability of wrong decisions with their implied losses.

In the general case of an arbitrary hypothesis test problem, let $L(\theta)$ denote the loss function as a random variable, and let $\psi\left(T(\tilde{\mathbf{X}})\right)$ denote a test function, i.e. $\psi\left(T(\tilde{\mathbf{X}})\right)$ returns 1 when H_0 is taken as true, but it returns 0 otherwise. For fixed θ, the expected loss, denoted by $r(\theta, \psi)$, is given by:

$$r(\theta, \psi) = E_\psi[L(\theta)] = l_1(\theta) \times Pr[\psi\left(T(\tilde{\mathbf{X}})\right) = 1|\theta] \times I_{\{\theta \in \Theta_0\}}(\theta) + \\ + l_2(\theta) \times Pr[\psi\left(T(\tilde{\mathbf{X}})\right) = 0|\theta] \times I_{\{\theta \in \Theta_1\}}(\theta). \qquad (5)$$

For example, if the Bayes Factor is the selected test statistic, then the expected loss associated to exact Bayesian test can be rewritten as:

$$r(\theta, \psi) = E_{BF}[L(\theta)] = l_1(\theta) \times Pr[BF(T(\tilde{\mathbf{X}})|\pi_\theta) \le c|\theta] \times I_{\{\theta \in \Theta_0\}}(\theta) + \\ + l_2(\theta) \times Pr[BF(T(\tilde{\mathbf{X}})|\pi_\theta) > c|\theta] \times I_{\{\theta \in \Theta_1\}}(\theta). \qquad (6)$$

3 Bayesian Monte Carlo Hypothesis Testing: our proposal

For simplicity, herein the expected loss of the exact Bayesian test is simply denoted by $r(\theta)$. Let $T(\tilde{\mathbf{X}})$ denote a real-valued statistic for an unknown parameter $\theta \in \mathbb{R}^N$. In certain problems, the use of a one-dimensional $T(\tilde{\mathbf{X}})$ is not a choice, but an inevitable path that the investigator has to follow due to particular characteristics of analysis. This is the case, for example, in intensive computational procedures appearing in the context of big data analysis. A practical example shall be offered in Section 3.1 for testing mean vectors in high dimension.

Without loss of generality, the reasoning is developed assuming that $T(\tilde{\mathbf{X}})$ has a continuous probability distribution, but all results can be easily extended to the discrete case by analogy. An application for real data analysis, in the context of spatial cluster detection, is offered in Section 5 to illustrate the method for the situation where data follows a multivariate discrete distribution.

The method introduced in the present section is valid for any of the following six pairs of hypotheses:

for $N = 1$,

$(i)\ \ H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$,

$(ii)\ \ H_0 : \theta \ge \theta_0$ versus $H_1 : \theta < \theta_0$,

$(iii)\ \ H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$,

$(iv)\ \ H_0 : \theta = \theta_0$ versus $H_1 : \theta \ne \theta_0$, $\qquad (7)$

and for $N > 1$,

$(v)\ \ H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$,

$(vi)\ \ H_0 : \theta = \theta_0$ versus $H_1 : \theta \ne \theta_0$. $\qquad (8)$

Let $T_1, \cdots, T_m$ denote a Monte Carlo sample of $T(\tilde{\mathbf{X}})$ somehow generated from $F_T(t|\theta = \theta_0)$, and use t_0 to denote the observed value of $T(\tilde{\mathbf{X}})$ for the actual observed sample $\tilde{\mathbf{x}}_0$, i.e., $t_0 = T(\tilde{\mathbf{x}}_0)$. Obviously, if nuisance parameters are present, the process of generating the Monte Carlo sample is supposed to be correctly adjusted for it. The details of how to generate a Monte Carlo sample of $T(\tilde{\mathbf{X}})$, and how to adjust it for nuisance parameters, depend of each problem. Later in this section a detailed descriptions of how to deal with nuisance parameters shall be offered for the problem of testing in high dimension.

Define the Monte Carlo measure of evidence, L_m, as follows:

$$L_m = \begin{cases} \sum_{i=1}^{m} I_{\{T_i \geq t_0\}}(T_i), & \text{if } T \text{ tends to be large under } H_1, \\ \sum_{i=1}^{m} I_{\{T_i \leq t_0\}}(T_i), & \text{if } T \text{ tends to be small under } H_1. \end{cases} \tag{9}$$

Let $r_m(\theta) > 0$ denote the expected loss of the Bayesian Monte Carlo test. Finally, we introduce the following Bayesian Monte Carlo decision rule: the null hypothesis is to be taken as true if $\psi = 0$, and the alternative hypothesis is taken as true if $\psi = 1$, where:

$$\psi = \begin{cases} 0, & \text{if } L_m > h_m, \\ 1, & \text{if } L_m \leq h_m, \end{cases} \tag{10}$$

$h_m = \lfloor (1+\varepsilon)(m+1)k/Pr[\theta \in \Theta_0|\pi_\theta] \rfloor - 1$, and $k = \{1 + c^{-1}(1 - Pr[\theta \in \Theta_0|\pi_\theta])/Pr[\theta \in \Theta_0|\pi_\theta]\}^{-1}$.

The term $\varepsilon > 0$ is a tuning parameter representing the arbitrary upper bound for the relative expected loss of the Monte Carlo test with respect to the exact test, i.e., the analyst can ensure that $r_m(\theta)/r(\theta) - 1 \leq \varepsilon$, where

$$r_m(\theta) = l_1(\theta)Pr[L_m \leq h_m|\theta]I_{\{\theta \in \Theta_0\}}(\theta) + l_2(\theta)Pr[L_m > h_m|\theta]I_{\{\theta \in \Theta_1\}}(\theta), \tag{11}$$

where, again, c is the critical value for the exact Bayesian test given in the scale of the Bayes factor. Observe that $r(\theta)$ and $r_m(\theta)$ cannot be calculated in practice because such computation requires a known $F_T(t|\theta)$. Thus, a quite important property of the decision rule in (10) is: there always exists a minimum value for m such that the bound ε is valid without requiring any knowledge about the values of $r(\theta)$ and $r_m(\theta)$. Surprisingly, this guaranteed control over the performance of the Monte Carlo test does not depend on the shape of the arbitrary loss function, $l_j(\theta)$.

Theorem 1 *Consider a hypothesis test problem having one of the forms (i) to (vi), and let $T(\tilde{\mathbf{X}})$ denote a real-valued sufficient statistic for θ. Use Θ_0 and Θ_1 to denote the subsets of the parameter space under the null and the alternative hypotheses, respectively. Use $r(\theta)$ to denote the expected loss of the exact Bayesian test constructed through the Bayes factor and based on $T(\tilde{\mathbf{X}})$. Let $r_m(\theta)$ denote expected loss of the Bayesian Monte Carlo test defined in (10). Thus, for $\varepsilon > 0$ arbitrary, there always exists $m_0 < \infty$ such that, for any $m \geq m_0$, it is valid that*

$$r_m(\theta)/r(\theta) - 1 \leq \varepsilon \text{ for each } \theta \in \Theta,$$

and moreover:

$$m_0 = \min\{m^* \in \mathbb{N}^+ : \sum_{g=h_{m^*}+1}^{m^*} C_g^{m^*} \gamma^g (1-\gamma)^{m^*-g} \leq \varepsilon\}, \tag{12}$$

where $C_g^m = m!/[g!(m-g)!]$, and $\gamma = (Pr[\theta \in \Theta_1|\pi_\theta]/c + Pr[\theta \in \Theta_0|\pi_\theta])^{-1}$.

Proof For a lighter notation, $T(\tilde{\mathbf{X}})$ shall be referred simply by T, and the representation of the prior probability of H_0, $Pr[\theta \in \Theta_0|\pi_\theta]$, is replaced by the shorter notation 'π_0'. The test based on the Bayes factor takes H_1 as true if $BF(t_0|\pi_\theta) \leq c$, where t_0 is an observed value of T. Define the event $A(t_0) = \{\mathbf{x} \in \aleph : T(\mathbf{x}) \leq t_0\}$, where $\aleph$ is the sample space of $\tilde{\mathbf{X}}$. Note that $BF(t_0|\pi_\theta) \leq c$ iff $BF(A(t_0)|\pi_\theta) \leq c$. Let Q denote the posterior probability of H_0 before observing a realization of T for a fixed π_0, videlicet $Q = Pr[\theta = \theta_0|A(T), \pi_\theta]$, and use q to denote a realization of Q after observing $T = t_0$. Thus, the probability of taking H_1 as true is $Pr[Q \leq k|\theta]$ calculated with respect to the distribution of T, where $k = [1 + (c\pi_0)^{-1}(1 - \pi_0)]^{-1}$. Therefore,

$$
\begin{aligned}
q &= Pr[\theta = \theta_0|A(t_0), \pi_\theta] \\
&= \int_{\Theta_0} Pr[A(t_0)|\theta = y]\pi_\theta(y)dy / Pr[A(t_0)] \\
&\geq \pi_0 Pr[A(t_0)|\theta = \theta_0]/Pr[A(t_0)] \\
&\geq \pi_0 Pr[A(t_0)|\theta = \theta_0] = \hat{p}.
\end{aligned}
$$

Now take the auxiliary random variable $\hat{P} = \pi_0 Pr[A(T)|\theta = \theta_0]$. Observe that, due to the inequalities above:

$$Pr[Q \leq k|\theta] \leq Pr[\hat{P} \leq k|\theta], \text{for each } \theta \in \Theta.$$

Consider the second auxiliary function $P = Pr[A(T)|\theta = \theta_0]$, and denote the probability densities of $\hat{P}$ and of P by $f_{\hat{P}}(\hat{p}|\theta)$ and $f_P(p|\theta)$, respectively. For the difference between the exact and the Monte Carlo tests probabilities of taking H_1 as true, denoted by $d(m,k)$, holds the following:

$$
\begin{aligned}
d(m,k) &= Pr[P \leq k|\theta] - Pr[L_m \leq h_m] \\
&\leq \int_0^k f_{\hat{P}}(\hat{p}|\theta)d\hat{p} - \int_0^1 f_P(p|\theta)Pr[L_m \leq h_m|P = p]dp.
\end{aligned}
$$

By noting that $\hat{p} = p \times \pi_0$, and after an adequate adjustment for the variation of the integrands, we have:

$$d(m,k) \leq \int_0^1 f_P(p|\theta) \left\{ I_{\{p \leq k/\pi_0\}}(p) - Pr[L_m \leq h_m|P = p] \right\} dp \tag{13}$$

Then, because $Pr[L_m \leq h_m|P = p]$ is decreasing with p, the integrand in (13) is maximized at $p = k/\pi_0$, i.e.,

$$argmax_{p \in (0,1)} \left\{ I_{\{p \leq k/\pi_0\}}(p) - Pr[L_m \leq h_m|P = p] \right\} = k/\pi_0,$$

impplying that:$d(m,k) \leq 1 - Pr[L_m \leq h_m|P = \gamma]$, where $\gamma = k/\pi_0$. Therefore,

$$
\begin{aligned}
&\frac{1 - Pr[L_m \leq h_m]}{1 - Pr[P \leq k]} \leq 1 - Pr[L_m \leq h_m|P = \gamma], \\
\Rightarrow &\frac{r_m(\theta)}{r(\theta)} = \frac{(1 - Pr[L_m \leq h_m])l_2(\theta)}{(1 - Pr[P \leq k])l_2(\theta)} \leq 1 - Pr[L_m \leq h_m|P = \gamma].
\end{aligned}
$$

The term $Pr[L_m \leq h_m|P = \gamma]$ is increasing with $(m+1)$ in the multiples of $\lfloor 1/\gamma \rfloor$. Thus, for $\varepsilon > 0$ arbitrary, there always exists m_0 such that, for $m \geq m_0$,

$$1 - Pr[L_m \leq h_m|P = \gamma] = \sum_{g=h_m+1}^{m} C_g^m \gamma^g (1-\gamma)^{m-g} \leq \varepsilon. \tag{14}$$

Now, it is trivial to note that $Pr[P \leq p|\theta = \theta_0] = p$, for $p \in (0,1)$, i.e., for $\theta = \theta_0$, the variable P follows an uniform probability distribution with support in the $(0,1)$ interval. Also, the maximum difference of the probabilities of taking H_0 as true between the exact Bayesian test and the Monte Carlo Bayesian test occurs for $\theta = \theta_0$. Thus, for $m \geq m_0$,

$$r_m(\theta)/r(\theta) \leq \frac{\int_0^1 Pr[L_m \leq h_m|P = p]dp}{\int_0^{k/\pi_0} dp} = \frac{\int_0^1 \sum_{g=0}^{h_m} C_g^m p^g (1-p)^{m-g} dp}{k/\pi_0}$$

$$\Rightarrow r_m(\theta)/r(\theta) \leq \frac{(h_m+1)/(m+1)}{k/\pi_0} = (1+\varepsilon). \tag{15}$$

In conclusion, situations where the Monte Carlo test can present increased expected losses in comparison to the exact test are properly bounded according to inequalities (14) and (15).

□

Theorem 1 reveals that the number of simulations required to ensure the performance of the Monte Carlo test does not depend on the shapes of the loss function and of the prior distribution, but it only depends on the amount of probability mass allocated to H_0 through the prior, and on the arbitrary critical value c. For example, suppose that a certain analyst does not admit an expected loss difference greater than 2% between the Monte Carlo and the exact test, i.e $\varepsilon = 0.02$. If the adopted critical value in the Bayes factor scale is $c = 0.01$, in the Jeffrey's scale the critical value is equal to $-\ln_{10} 0.01 = 2$. Also, suppose that the analyst's uncertainty sets up $\pi_0 = Pr[\theta \leq \theta_0|\pi_\theta] = 0.1$. Then, $k = 1/[1+(1-0.1)/(0.1 \times 0.01)] \approx 0.00111$, with $\gamma = 0.00111/0.1 = 0.0111$ and $h_m = \lfloor (1+0.02) \times (m+1) \times 0.00111/0.1) \rfloor - 1$. Solving inequality (12) for m, the respective lower bound for the number of Monte Carlo simulations complying with $\varepsilon = 0.02$ is $m_0 = 949,230$, which gives $h_m = 10,745$. Naturally, each combination of ε and π_0 gives a different m_0 solution.

Table 1 Minimum number of Monte Carlo Simulations for an upper bound on the relative loss of $\varepsilon = 0.01, 0.02, 0.03, 0.04, 0.05, 0.1$, prior probabilities for the null hypothesis of $\pi_0 = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$, and critical value in the Bayes Factor scale equal to $c = 0.01$.

		ε					
		0.01	0.02	0.03	0.04	0.05	0.1
$\pi_0 = 0.05$	m_0	5116795	1005480	377340	185440	105735	17195
	h_m	54370	10789	4088	2028	1167	198
$\pi_0 = 0.1$	m_0	4845060	949230	358650	176130	99540	16470
	h_m	54311	10745	4099	2032	1159	200
$\pi_0 = 0.25$	m_0	4040775	794625	299100	145725	83850	13275
	h_m	54234	10770	4093	2013	1169	193
$\pi_0 = 0.5$	m_0	2691050	530150	199450	97600	55550	9000
	h_m	53820	10707	4067	2009	1154	195
$\pi_0 = 0.75$	m_0	1345450	264400	98800	48825	28275	4400
	h_m	52772	10472	3951	1971	1152	187
$\pi_0 = 0.9$	m_0	538060	105420	39590	19400	11120	1910
	h_m	49856	9864	3740	1850	1070	191
$\pi_0 = 0.95$	m_0	268945	52645	19715	9685	5490	870
	h_m	45652	9024	3412	1692	967	160

Table 1 contains m_0 solutions for $\pi_0 = 0.05, 0.1, 0.25, 0.5, 0.75, 0.9, 0.95$ and $\varepsilon = 0.01, 0.02, 0.05, 0.1$. The critical values in the scale of L_m, h_m, are presented in this table too.

Observe that m_0 is a decreasing function of π_0 for fixed ε. Analogously, when π_0 is held fixed, m_0 is decreasing with ε.

Although the Bayesian practice has been largely diffused as being limited to the particular and trivial choice $T(\tilde{\mathbf{X}}) := \tilde{\mathbf{X}}$, there is also a number of problems that are well solved under different and more elaborated sufficient functions. Supported by the well-established data reduction principle, a real-valued sufficient statistic can be used to solve interesting multivariate problems. This is the case, for example, of the test statistic proposed by Nyblom and Makelainen (1983) for testing random effects in state space models, of the cMaxSPRT statistic for sequential analysis of adverse events with Poisson data (Li and Kulldorff, 2010), and of the scan statistic for detection of spatial clusters Kulldorff and Nagarwalla (1995). An important problem solved through usage of a real-valued sufficient statistics is testing mean vectors in high dimension, which is described in details in what follows.

3.1 Testing Mean Vectors in High Dimension

A classical multivariate method for testing the mean vector, μ, of a P-variate normal observation $\tilde{X}_j = (X_{j,1}, \cdots, X_{j,P})$, $j = 1, \cdots, n$, is the Hotelling's test, which is based on the following sufficient statistic for μ:

$$T^2 = n(\bar{X} - \mu_0)'S^{-1}(\bar{X} - \mu_0), \tag{16}$$

where n is the sample size, $\bar{X} = \sum_{j=1}^{n} \tilde{X}_j/n$, and $S = \frac{1}{n}\sum_{j=1}^{n}(\tilde{X}_j - \mu_0)(\tilde{X}_j - \mu_0)'$. It is possible to prove that T^2 is an ancillary statistic for Σ, the unknown covariance matrix of $\tilde{X}_j$. While S is an unbiased estimator of Σ, it is also singular when $n < P$. In practice, data can lead to a high dimensional situation, i.e. n may be much smaller than P and so T^2 cannot be calculated since S becomes singular. In order to obtain an invertible estimate of Σ, Maboudou-Tchao and Silva (2013) and Silva et al (2018) suggest to apply the so called 'slicing' estimator, which by its turn was introduced by Akdemir and Gupta (2011). The idea is decomposing Σ in the Kronecker product of two auxiliary matrices, the p-dimensional Ω and the q-dimensional Ψ, that is $\Sigma = \Omega \otimes \Psi$, where $P = p \times q$. Consequently, the maximum likelihood estimator of Σ^{-1} given by

$$\hat{\Sigma}_\otimes^{-1}(\tilde{\mathbf{X}}) = \hat{\Omega}^{-1}(\tilde{\mathbf{X}}) \otimes \hat{\Psi}^{-1}(\tilde{\mathbf{X}})$$

can be used to construct the sufficient statistic $T_s^2(\tilde{\mathbf{X}})$, obtained by replacing S^{-1} by $\hat{\Sigma}_\otimes^{-1}(\tilde{\mathbf{X}})$ in (16):

$$T^2 = n(\bar{X} - \mu_0)'\hat{\Sigma}_\otimes^{-1}(\tilde{\mathbf{X}})(\bar{X} - \mu_0), \tag{17}$$

where $\tilde{\mathbf{X}}$ is the $(n \times P)$-dimensional matrix formed by the sample of n observations. Although the distribution of $T_s^2(\tilde{\mathbf{X}})$ is unknown, Maboudou-Tchao and Silva (2013) showed that: (i) $T_s^2(\tilde{\mathbf{X}})$ is ancillary for Σ, hence the expected loss associated to the Bayes factor having $T_s^2(\tilde{\mathbf{X}})$ as the empirical information is invariant with respect to the actual Σ; and (ii) for arbitrary μ^*, Monte Carlo samples of $T_s^2(\tilde{\mathbf{X}})$ can be obtained by applying independent observations $\tilde{Z}_j$, $j = 1, \cdots, n$, in the function $T_s^2(\tilde{Z})$, where $\tilde{Z}_j \sim N_P(\mu^*, I_P)$. Hence, the Bayesian Monte Carlo test can be naturally applied for arbitrary prior multivariate distributions for μ without requiring a prior distribution for the nuisance matrix, Σ.

4 Comparison Against Competing Methods

The performance of ABC methods depends on assumptions about the shape of the likelihood and on asymptotic evaluations. Moreover, there is no general rules for the choice of the tuning parameters called 'discrepancy' and 'tolerance' that are to be informed when an ABC procedure is implemented. The best or worse tuning parametrization setting depends on the shape of the likelihood, which is then a serious limitation when the likelihood shape is unknown. Therefore, the main advantage of the Bayesian Monte Carlo test is its simplicity and generality. As emphasized with Theorem 1, the tuning parametrization does not depend on the likelihood shape, but solely on a logical, simple, and general rule. Moreover, the method is valid for any sample size, i.e. its performance is not limited to asymptotic properties.

But, one could question if ABC would be preferable than the Bayesian Monte Carlo test when tuning parametrization of ABC is also simple and safe to use. The simplicity of the Bayesian Monte Carlo test should not penalize its performance in both, simple and complex problems. Thus, the question arising is: is there some loss if we use the Bayesian Monte Carlo test instead of conventional methods? Aiming to answer this question, in this section we use a very simple Poisson testing problem to compare the Bayesian Monte Carlo test (*BMCT*) against the Importance Sampling (*IS*), the Move-Reweigthing (*MR*), and the Single Auxiliary Variable Importance Sampling (*SAVIS*) algorithms (see (Marques and Storvik, 2013) and (Didelot et al, 2011) for description of the algorithms). Although *BMCT* is theoretically valid for the general case of any $F_T(t|\theta)$ shape, it is important to clarify if it performs well, comparatively to competing methods, for simple and classical testing problems too.

Note that it is not fair to consider the Importance Sampling method in this comparison study because it requires knowledge about the likelihood shape, and hence, unlike the other three methods, *IS* is privileged with additional information that, in practice, is not available when the shape of $F_T(t|\theta)$ is unknown but *BMCT*, *MR* and *SAVIS* are feasible to use. But, *IS* is included in this study to show that, even though *BMCT* requires much less information to use, it is even comparable to a method such as *IS* that requires much information such as knowledge about the likelihood shape.

Let $X_1, \cdots, X_n$ denote a random sample from a Poisson distribution with unknown parameter θ. The hypotheses of interest are:

$$H_0 : \theta \geq 4 \text{ versus } H_1 : \theta < 4. \tag{18}$$

The choice $\theta_0 = 4$ serves merely to illustrate the type of results found if one uses this simple and naive problem involving the Poisson distribution when comparing these methods. Actually, we have reached the same conclusions (results not shown here) after running simulations for $\theta_0 = 1, 2, \cdots, 10$.

Here we consider the simple conjugate prior distribution for Poisson data, the gamma function:

$$\pi_\theta(y) = \frac{ye^{-2y}}{2}I(y > 0). \tag{19}$$

The statistical performance measure of interest in this comparison study is the expected loss. For this, we use the following loss function:

$$\begin{cases} l_1(\theta) = (\theta_0 - \theta)^2 I(\theta \geq \theta_0), \\ l_2(\theta) = (\theta_0 - \theta) I(\theta < \theta_0). \end{cases} \tag{20}$$

This loss function is simply for illustrative matters, hence here it does not represent any practical unit in particular. According to Casella and Berger (2001), page 401, expression

Table 2 Comparison of expected loss among Bayesian Monte Carlo test (*BMCT*), Importance Sampling (*IS*), Move-Reweigthing (*MR*), and Single Auxiliary Variable Importance Sampling (*SAVIS*). Each method was settled to return zero loss at $\theta = \theta_0 = 4$.

		BMCT	IS	MR	SAVIS
	$\theta = 3.0$	0.50060	0.51790	0.50890	0.47600
	$\theta = 3.2$	0.55760	0.56992	0.56328	0.54272
	$\theta = 3.4$	0.49932	0.50418	0.50274	0.49182
	$\theta = 3.6$	0.37128	0.37312	0.37208	0.36812
	$\theta = 3.8$	0.19454	0.19492	0.19470	0.19400
$n = 20$	$\theta = 4.0$	0	0	0	0
	$\theta = 4.2$	0.00015	0.00012	0.00014	0.00017
	$\theta = 4.4$	0.00013	0.00011	0.00011	0.00013
	$\theta = 4.6$	0.00011	0.00011	0.00011	0.00007
	$\theta = 4.8$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$
	$\theta = 5.0$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$
	$\theta = 3.0$	0.07300	0.08140	0.08210	0.07820
	$\theta = 3.2$	0.21296	0.22184	0.22424	0.22064
	$\theta = 3.4$	0.33276	0.34002	0.34308	0.34038
	$\theta = 3.6$	0.32784	0.33000	0.33120	0.33124
	$\theta = 3.8$	0.18964	0.18982	0.18996	0.19010
$n = 50$	$\theta = 4.0$	0	0	0	0
	$\theta = 4.2$	0.00007	0.00006	0.00005	0.00006
	$\theta = 4.4$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$
	$\theta = 4.6$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$
	$\theta = 4.8$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$
	$\theta = 5.0$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$	$< 10^{-6}$

(8.3.13), this loss function shape was advocated by Vardeman (1987), and hence we just follow it since it is well-accepted in the classical theory on loss functions.

The expected loss was evaluated for $\theta = 3, 3.2, \cdots, 5$, for moderate ($n = 20$) and regular ($n = 50$) sample sizes. For each of these settings, the expected loss was estimated on a total of 10,000 simulated n-dimensional samples. Also, each of the four methods were run for $m = 10,000$. It is important to point out that the critical value used for each method was settled in such a way that all of them returned with zero loss under $\theta = \theta_0$.

Table 2 contains the main results of this simulation study. In general, the *BMCT* method presents expected losses very close or even better than *IS*. This is curious since that the tuning parametrization of *BMCT* does not require any knowledge about the likelihood shape and, unlike ABC based approaches, there is a general and simple rule for setting the tuning parametrization of *BMCT*, which follows from Theorem 1.

5 Real Data Analysis - spatial cluster detection

It is quite common to define the Bayes Factor for the particular statistic $T(\tilde{\mathbf{X}}) := \tilde{\mathbf{X}}$, which leads to the likelihood corresponding to the product of the densities of the $X_i's$ and evaluated at the observed sample, $\tilde{\mathbf{x}}$. But, it is also possible to face problems where the interest is to use a different and more elaborated function of the data for which the likelihood cannot be expressed analytically. This is the case of the circular Scan statistic. The Scan statistic is extensively used in the detection of spatial clusters. According to Cressie (1993), a spatial cluster is a set of non-overlapping neighboring areas, indexed from 1 to K, of a region (map) with elevated risk for the occurrence of a certain event of interest.

For a given radium and centroid, let z to be the set of all areas with centroids inside the circle. Let C_i, $i = 1, \cdots, B$, be the number of events inside a given circle z_i. Suppose that C_i follows a Poisson distribution with expectation equal to $\theta_i \lambda N_i$, where N_i is the population inside z_i, λ is a base expectation associated to the nature of the problem, and θ_i is the relative risk of occurring a case inside z_i. If the possible increased risk is assumed to be under control for all areas inside a cluster, the hypotheses to be tested are $H_0 : \theta \leq 1$, for all $i = 1, \cdots, B$, versus $H_a : \theta > 1$ for some i in $[1, B]$. The test statistic proposed by Kulldorff and Nagarwalla (1995) is defined as the maximum, denoted by Λ, among all the log-likelihood ratio (LLR) statistics. Using N to denote the total population size of the map, the log-likelihood ratio statistic for the circle z_i is equal to:

$$
\begin{aligned}
LLR(z_i) = {} & \ln \left[\left(\frac{C_i}{N_i} \right)^{C_i} \left(\frac{N_i - C_i}{N_i} \right)^{N_i - C_i} \right] + \\
& + \ln \left[\left(\frac{C - C_i}{N - N_i} \right)^{C - C_i} \left(\frac{(N - N_i) - (C - C_i)}{N - N_i} \right)^{(N - N_i) - (C - C_i)} \right] - \\
& - \ln \left[C^C (N - C)^{N - C} / N^N \right],
\end{aligned}
\tag{21}
$$

if $\frac{C_i}{N_i} > \frac{C - C_i}{N - N_i}$, and $LLR(z_i)$ is equal to zero, otherwise.

The Scan statistic is a powerful tool for detection and identification of spatial clusters. Thus, the use of the Scan statistic in this original form is desired. But, the exact Bayesian hypothesis test based on $T(\tilde{\mathbf{X}}) := \Lambda$ is not viable because the exact distribution of Λ is intractable even for small maps, then the likelihood with respect to Λ is unknown. Fortunately, the Monte Carlo test is feasible. Conditioning in the total number (C) of events observed in the map, the joint distribution of the numbers of events in the areas is multinomial, where the probability of having a event inside area k is equal to $\theta_k N_k / (\sum_{l=1}^{K} \theta_l N_l)$. Thus, a sample of Λ, under $\theta = \theta_0 = 1$, can be generated by simulating events in the map from a multinomial distribution.

5.1 Data Description

The Bayesian Monte Carlo test is illustrated for the extensively analyzed database of brain cancer incidence in New Mexico, United States ('http//www.satscan.org/datasets'). Beside the spatial analysis, a temporal exploration is also possible because the data was collected for the period 1973-1991. But, for simplicity, the cases of each area were concatenated in just a single count. The map consists of 33 counties. In the period 1973-1991, it was registered a total of 1,175 cases of brain cancer, and the total exposed population in the map at 1991 was 1,548,640. The geographical centroid location (latitude/longitude) of each county is also available in the same site. A detailed description and full analysis for this data set is offered by Kulldorff et al (1998). The map showed in the top of Figure 1 presents the observed relative risks for each county.

5.2 Tuning Parameters Settings

For the prior distribution, assume that $Pr[\theta \leq 1 | \pi_\theta] = Pr[\theta > 1 | \pi_\theta] = 0.5$. Concerning the relative performance tolerance, set ε equal to 0.02. Following Jeffrey's scale, the critical value associated to the exact test in the Bayes Factor scale is settled at $c = 0.01$ (it corresponds to a value of 2 in the minus-log scale). These settings lead to $k = 0.0099$ and

$\gamma = 0.0099/0.5 = 0.0198$. Using Theorem 1, and solving inequality (12) for m, the solution is $m_0 = 530,150$ Monte Carlo simulations. Finally, the resulting critical value in the scale of the Monte Carlo measure of evidence is $h_m = 10,707$.
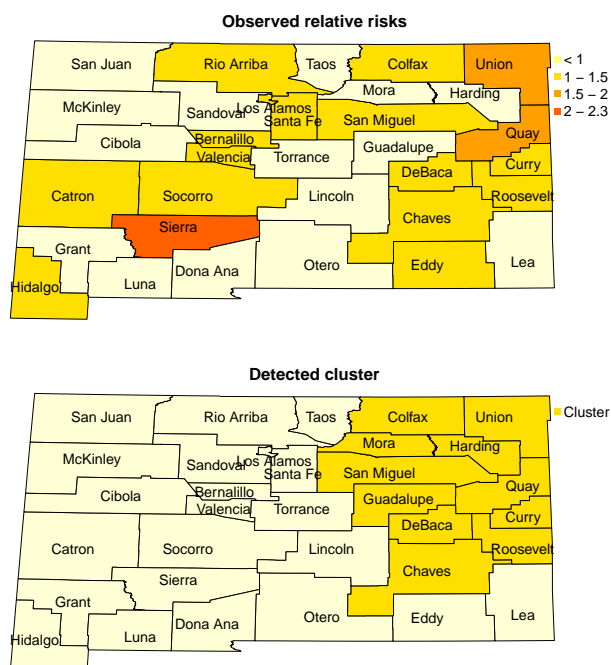

Fig. 1 Relative risk estimates for each county (top part) and for the detected cluster (bottom part).

For this data, the estimated relative risk was $\hat{r} = 1.296$, and the observed value of Λ was $t_0 = 5.4587$. The observed cluster is formed by the following counties: Chaves, Colfax, Curry, DeBaca, Guadalupe, Harding, Mora, Quay, Roosevelt, San Miguel, and Union. This detected cluster is highlighted with a different color in the map located in the bottom part of Figure 1. With a population of 185,694 individuals, this zone gathered 179 cases. After generating $530,150$ values of Λ under $\theta = 1$, the observed value of L_m was 1,937, which is smaller than the critical value $h_m = 10,707$. Thus, the evidence in favor of H_1 is strong enough for taking it as true according to the Bayesian Monte Carlo test. In conclusion, this analysis reveals the existence of a spatial cluster in this map for the considered period.

6 Concluding Remarks

Conventional approaches for dealing with Bayesian testing through Monte Carlo are funded in the spirit of integrating the likelihood function with respect to the prior distribution. Such approaches can be interpreted as indirect ways to approximate the entire posterior distribution in order to construct an estimate of the adopted measure of evidence (test statistic). Although intuitive, the idea of estimating the whole posterior distribution for approximating the test statistic is not mandatory neither the only way for treating the cases where Monte Carlo simulation is needed. This paper shows that the usual practice of generating observations from the prior distribution (in the spirit of integrating the likelihood with respect to the prior) can be discarded in hypothesis testing. This result contradicts the belief among some that a legitimate Bayesian Monte Carlo procedure must involve, somehow, samples generated from the prior distribution. The ultimate goal of a Bayesian Monte Carlo test is not to figure out a way to estimate the posterior distribution. Unlike, for a given prior distribution reflecting the investigator's uncertainty about θ, the ultimate goal of a Monte Carlo testing procedure is to provide statistical performance arbitrarily close to the exact Bayesian test.

With the procedure introduced in this paper, the true management of the relative expected losses implied by the Bayesian Monte Carlo test is proved for finite sample sizes under arbitrary prior distributions. Besides the simplicity, the Bayesian Monte Carlo test follows the strict Bayesian reasoning in which concerns usage of the Bayes Factor for drawing a decision, and it is virtually general in the sense of being valid for any unknown likelihood shape. Therefore, when the likelihood function is unknown, and if simulating the test statistic is a feasible option, there is no reason to deliberately base the hypothesis test decision on methods that require specific prior distribution shapes elaborated to accomplish with mathematical conveniences. Hence, the method proposed in this paper ensures that the analyst's free-will for using its actual prior uncertainty about θ can be preserved when using Monte Carlo methods for testing.

References

Akdemir D, Gupta AK (2011) Array variate random variables with multiway kronecker delta covariance matrix structure. J Algebr Stat 2(1):98113

Andrieu C, Doucet A, Holenstein R (2010) Particle markov chain monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72(3):269–342

Besag J (2001) Markov chain monte carlo for statistical inference. Working Paper n9 pp 1–67

Casella G, Berger RL (2001) In: Statistical Inference - Second Edition, Duxbury Resource Center, Thomson Learning

Chen MH, Shao QM, Ibrahim JG (2012) Monte Carlo methods in Bayesian computation. Springer Science & Business Media

Cressie N (1993) In: Statistics for Spatial Data, John Wiley and Sons, Inc

Didelot X, Everitt RG, Johansen AM, Lawson DJ, et al (2011) Likelihood-free estimation of model evidence. Bayesian analysis 6(1):49–76

Everitt RG, Johansen AM, Rowing E, Evdemon-Hogan M (2017) Bayesian model comparison with un-normalised likelihoods. Statistics and Computing 27(2):403–422

Fearnhead P (2017) Asymptotics of abc. arXiv preprint arXiv:170607712

Gilks WR, Berzuini C (2001) Following a moving targetnte carlo inference for dynamic bayesian models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63(1):127–146

Jeffrey H (1961) In: The Theory of Probability (3 ed.) p.432, Oxford

Kulldorff M, Nagarwalla N (1995) Spatial disease clusters: Detection and inference. Statistics in Medicine 14:799–810

Kulldorff M, Athas W, Feuer E, Miller B, Key C (1998) Evaluating cluster alarms: A space-time scan statistic and brain cancer in los alamos. American Journal of Public Health 88:1377–1380

Li L, Kulldorff M (2010) A conditional maximized sequential probability ratio test for pharmacovigilance. Statistics in Medice 29:284–295

Maboudou-Tchao EM, Silva IR (2013) Tests for mean vectors in high dimension. Statistical Analysis and Data Mining 6(6):578–598

Marin JM, Pillai N, Robert C, Rousseau J (2014) Relevant statistics for bayesian model choice. Journal of the Royal Statistical Society, Series B 76(5):833–859

Marques R, Storvik G (2013) Particle move-reweighting strategies for online inference. Statistical Research Report Matematisk Institutt, Universitetet i Oslo (1)

Nyblom J, Makelainen T (1983) Comparisons of tests for the presence of random walk coefficients in a simple linear model. Journalof the American Statistical Association 78(384):856–864

Owen AB, Zhou Y (2000) Safe and effective importance sampling. Journal of the American Statistical Association 95(449):1353

Pereira CAB, Stern JM (1999) Evidence and credibility: full bayesian significance test of precise hypothesis. Entropy 1:99–110

Price LF, Drovandi CC, Lee A, Nott DJ (2017) Bayesian synthetic likelihood. Journal of Computational and Graphical Statistics pp 1–11

Robert C, Casella G (2013) Monte Carlo statistical methods. Springer Science & Business Media

Silva IR, Maboudou-Tchao EM, Figueiredo WD (2018) Frequentistyesian monte carlo test for mean vectors in high dimension. Journal of Computational and Applied Mathematics (333):51–64

Sisson SA, Fan Y, Tanaka MM (2007) Sequential monte carlo without likelihoods. Proceedings of the National Academy of Sciences 104(6):1760–1765

Vardeman SB (1987) Discussion of the articles by casella and berger and berger and sellke. Journal of the American Statistical Society 82:130–131