Approximate Bayesian Inferences: Numerical experiences when measuring the loss of information

Michel MOUCHART[*] and Eliana SCHEIHING [*, †]

December 1994

Abstract

In this work, we first analyse the general problem of admissible conditioning and next consider the evaluation of the loss of information when a non-admissible conditioning is used for an approximation of the exact posterior distribution. Numerical experiences are presented in the particular case of the 2x2 contingency table when inference is based on the posterior distribution conditionally on the margins. More specifically, we compare two ways of measuring that loss of information, namely by means of a divergence between the approximate and the exact posterior distributions of the parameter of interest and by means of a relative efficiency between the risk functions based respectively on the exact and on the approximate posterior distributions.

KEY WORDS: Fisher Test, Bayesian Approximation, Bayesian Conditioning, Information Loss
Acknowledgement : Financial support from the Communauté Française de Belgique (Projet ARC N°93/98-164) is gratefully acknowledged.

[*] Institut de Statistique, Université Catholique de Louvain (Belgium)
[†] Instituto de Informática, Universidad Austral de Chile

1 Introduction

In this paper we develop methods to evaluate, from a Bayesian point of view, the loss of information when a sampling process is specified only partially by treating part of the data "as if" it were not random. We utilize the Fisher test, i.e. a test conditional on the two margins in a 2x2 contingency table, as a case study to quantify the loss of information by means of a simulation analysis.

Admissible Conditioning

Let a Bayesian model be a (unique) probability measure jointly on the parameters θ and the observations x:

$$q(x, \theta) = m(\theta)\, p(x \mid \theta) = p(x)\, m(\theta \mid x) \tag{1}$$

We decompose both the observations x and the parameter θ into two components, i.e. $x = (y, z)$ and $\theta = (\psi, \lambda)$, where ψ is a nuisance parameter and λ is a parameter of interest. The posterior distribution of λ may accordingly be decomposed as follows:

$$m(\lambda \mid x) \quad \propto \quad m(\lambda)\, p(z \mid \lambda)\, p(y \mid z, \lambda) \tag{2}$$

If z and λ are mutually ancillary (i.e. $z \perp\!\!\!\perp \lambda$) we also have:

$$m(\lambda \mid x) \quad \propto \quad m(\lambda)\, p(y \mid z, \lambda) \tag{3}$$

making the reduction by conditioning on z admissible. For a more systematic exposition see e.g. Florens et al (1990), chapter 3.

Measuring the loss of information

Suppose now that z and λ are not mutually ancillary but that one decides, for some reason, to "forget" $p(z \mid \lambda)$ in equation (2), i.e. using (3) for such an "approximation" of (2). Thus let us write $m_A(\lambda \mid x)$, as an approximation to $m(\lambda \mid x)$, i.e.:

$$m_A(\lambda \mid x) \propto m(\lambda)\, p(y \mid z, \lambda) \tag{4}$$

without having $z \perp\!\!\!\perp \lambda$. A natural question is "how much information do we lose by overlooking the term $p(z \mid \lambda)$?". We consider two ways of evaluating this loss of information. The first one is:

$$L = E[D(m_A(\lambda \mid x),\, m(\lambda \mid x))] \tag{5}$$

where $D(\ ,\ )$ is a divergence (or a distance) between probability measures and the expectation is taken with respect to exact predictive distribution. The second way compares the risk functions. Under the loss function $l(\lambda, a)$, the optimal decision rules under the exact (E) and under the approximate (A) posterior distribution of λ are:

$$\begin{aligned} a_E^*(x) &= arg \inf_a E_{\lambda|x}\, l(\lambda, a) &(6)\\ a_A^*(x) &= arg \inf_a E_{\lambda|x}^A\, l(\lambda, a) &(7)\end{aligned}$$

where

$$E_{\lambda|x}^A l(\lambda, a) = \int l(\lambda, a)\, m_A(\lambda \mid x)\, d\lambda$$

and the corresponding risk functions are:

$$\begin{aligned} \rho(a_E^*) &= E_{x,\lambda} l(\lambda, a_E^*(x)) &(8)\\ \rho(a_A^*) &= E_{x,\lambda} l(\lambda, a_A^*(x)) &(9)\end{aligned}$$

where the expectation is taken in both cases with respect to the exact joint distribution of (x, λ). Therefore we define the "Relative Efficiency of the approximation m_A" as:

$$RelEff(m_A) = \frac{\rho(a_E^*)}{\rho(a_A^*)} \in [0, 1] \tag{10}$$

As a particular case, consider the square loss function: $l(\lambda, a) = (a - \lambda)^2$, then

$$RelEff(m_A) = \left(1 + \frac{E_x[(\,E(\lambda \mid x) - E^A(\lambda \mid x)\,)^2]}{E_x\, V(\lambda \mid x)}\right)^{-1} \tag{11}$$

which shows that under a quadratic loss function, an approximation will be evaluated according to the ratio between the (predictively) expected exact posterior variance and the (predictively) expected square bias of the approximate posterior expectation.

The case of Fisher Test

Turning to the 2x2 contingency table, it is shown in Mouchart and Scheihing (1993) that under multinomial or independent binomial samplings, conditioning on the margins is generally not an admissible reduction. Then, we want to evaluate numerically how much information is lost when the inference on λ, the usual cross product ratio, is based on the sampling process conditional on the margins rather than on the actual sampling process. In particular we want to evaluate the respective roles of the sample size and the relative precision of the prior information.

The results show that the numerical problems raised by the comparison of the risk functions are substantially more complicated than those involved in the comparison of posterior distributions. Furthermore, all these measures of loss of information behave more or less equivalently as regards the role of the sample size. Namely, the approximation deteriorates as the sample size increases. This approximation is therefore not justified on asymptotic grounds. The role of the relative precision of the prior information, n_0, is more complex. For the discrepancy measures, the approximation clearly improves when n_0 increases. For the relative efficiency a non monotone behavior is found. In fact, the relative efficiency involves a ratio between two expected values (see equation (11)), both tending to 0 when n_0 goes to infinity, but at different rates. Finally, in both cases, the approximation can be qualified as a "small sample approximation" in the sense that the error of approximation tends to vanish when the prior precision tends to dominate the sample precision.

In the next section we describe the simulation design, while in section 3 we present the numerical results along with an analysis of the simulation accuracy.

2 Simulation Design

Firstly, we want to evaluate an expected discrepancy between the posterior distribution of parameter λ, $m(\lambda \mid x)$ and its approximation $m_A(\lambda \mid x)$, as written in (5). We consider two specifications for $D(\cdot,\cdot)$. The first one is the entropy divergence defined by:

$$\begin{aligned} D_E(m_A(\lambda \mid x),\, m(\lambda \mid x))] &= \int ln(\frac{m(\lambda \mid x)}{m_A(\lambda \mid x)})\, m(\lambda \mid x) d\lambda \qquad (12) \\ &\in (0,\infty) \end{aligned}$$

The second one is the Hellinger distance:

$$\begin{aligned} D_H(m_A(\lambda \mid x),\, m(\lambda \mid x))] &= \int (\sqrt{m(\lambda \mid x)} - \sqrt{m_A(\lambda \mid x)})^2\, d\lambda \\ &= 2(1 - \int \sqrt{m(\lambda \mid x)\, m_A(\lambda \mid x)}\, d\lambda) \qquad (13) \\ &\in [0,2] \end{aligned}$$

Sampling	A priori distribution	Sample size n	Weight of prior information n_0	Simulation size Discrep. measures	Rel. Eff.
Multinomial $(X \mid X_{..} = n, \theta) \sim MN(n, \theta)$	$\theta \sim D(n_0, p_0)$ $p_0 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	$(10, 15, 20, 25, 30, 35, 40, 45, 50)$	$(10, 15, 20, 25)$ $(30, 35, 40, 45, 50)$	1000 1000 drawings of (θ, X)	10000 5000
I. Binomial $(X_{1j} \mid X_{.j} = n, \beta_j) \sim I.Bi(n, \beta_j)$	$(\beta_1, \beta_2) \sim I.\beta(n_0, p_1)$ $p_1 = (\frac{1}{2}, \frac{1}{2})$	$(5, 10, 15, 20, 25, 30, 40, 50)$	$(5, 10, 15, 20, 25, 30, 40, 50)$	1000 drawings of $(\beta_j, X_{1j})_{j=1,2}$	5000

Table1: Simulation design for the experiments measuring the loss of information.

Secondly, we evaluate the Relative Efficiency for the square loss function, as written in (11) or equivalently:

$$RelEff(m_A) = \left(1 + \frac{E_x[(\, E(\lambda \mid x) - E^A(\lambda \mid x)\,)^2]}{E(\lambda^2) - E_x[(\, E(\lambda \mid x)\,)^2]}\right)^{-1} \tag{14}$$

For all these measures of comparison, we consider their behavior in relation to the values of two parameters: n the sample size and n_0 the relative precision of the prior information. In these experiments we simulate both cases of a multinomial sampling with Dirichlet prior and two independent binomial samplings with two independent Beta priors. Table 1 shows for each type of sampling the location parameter of the prior distribution, the simulated sample sizes and prior precisions and the simulation sizes for each measure of comparison.

In the multinomial case, $X \sim MN(n, \theta)$, the parameter of interest is defined as the usual cross product ratio:

$$\lambda = \frac{\theta_{11}\theta_{22}}{\theta_{12}\theta_{21}} \ , \tag{15}$$

the value 1 of which characterizes the row-column independence. When θ is distributed a priori as a Dirichlet distribution with parameter $n_0 p_0 = (a_{ij})$, it may be shown that:

$$m(\lambda|x) = \frac{\lambda^{b_4-1} I_1(\lambda, x)}{\int_0^\infty \lambda^{b_4-1} I_1(\lambda, x) d\lambda} \tag{16}$$

where

$$\begin{aligned} I_1(\lambda, x) &= \int_0^1 \delta^{b_1-1}(1-\delta)^{b_2-1}(1-\delta+\lambda\delta)^{-b_3} d\delta \\ b_1 &= x_{.1} + a_{.1} \\ b_2 &= x_{.2} + a_{.2} \\ b_3 &= x_{1.} + a_{1.} \\ b_4 &= x_{11} + a_{11} \end{aligned}$$

and

$$m_A(\lambda|x) = \frac{\lambda^{a_{11}-1} I_2(\lambda)\, p(x_{11}|x_{1.}, x_{.1}, \lambda)}{\int_0^\infty \lambda^{a_{11}-1} I_2(\lambda)\, p(x_{11}|x_{1.}, x_{.1}, \lambda) d\lambda} \tag{17}$$

where

$$p(x_{11}|x_{1.}, x_{.1}, \lambda) = \frac{\lambda^{x_{11}}}{\prod_{i,j} x_{ij}!} \left(\sum_{x_{11} \in V} \frac{\lambda^{x_{11}}}{\prod_{i,j} x_{ij}!} \right)^{-1} \tag{18}$$

$$V = \{\max\{0, x_{1.} + x_{.1} - x_{..}\}, ..., \min\{x_{1.}, x_{.1}\}\} \subset I\!N$$

$$\begin{aligned} I_2(\lambda) &= \int_0^1 \delta^{c_1-1}(1-\delta)^{c_2-1}(1-\delta+\lambda\delta)^{-c_3} d\delta \\ c_1 &= a_{.1} \\ c_2 &= a_{.2} \\ c_3 &= a_{1.} \end{aligned}$$

Furthermore, in the two independent binomial sampling, $(X_{1j} \mid X_{.j}, X_{..}, \beta) \sim Bi(X_{.j}, \beta_j)$, $j = 1, 2$, the parameter of interest is defined as:

$$\lambda = \frac{\beta_1(1-\beta_2)}{\beta_2(1-\beta_1)} \tag{19}$$

The value $\lambda = 1$, meaning $\beta_1 = \beta_2$, is often read as a condition of "homogeneity" and implies, again, a row-column independence of the 2x2 table for whatever process generating the column totals. When β_j, $(j = 1, 2)$, are distributed a priori as Beta distributions with parameters $n_0 p_1 = (a_j, b_j)$, $(j = 1, 2)$, it may be shown that:

$$m(\lambda|x) = \frac{\lambda^{d_4 - 1} I_3(\lambda, x)}{\int_0^\infty \lambda^{d_4-1} I_3(\lambda, x) d\lambda} \tag{20}$$

where

$$\begin{aligned}
I_3(\lambda, x) &= \int_0^1 \delta^{d_1-1}(1-\delta)^{d_2-1}(1-\delta+\lambda\delta)^{-d_3} d\delta \\
d_1 &= x_{1.} + a_{.} \\
d_2 &= x_{2.} + b_{.} \\
d_3 &= X_{.1} + a_1 + b_1 \\
d_4 &= x_{11} + a_1
\end{aligned}$$

and

$$m_A(\lambda|x) = \frac{\lambda^{a_1-1} I_4(\lambda)\, p(x_{11}|x_{1.}, x_{.1}, \lambda)}{\int_0^\infty \lambda^{a_1-1} I_4(\lambda)\, p(x_{11}|x_{1.}, x_{.1}, \lambda) d\lambda} \tag{21}$$

where $p(x_{11}|x_{1.}, x_{.1}, \lambda)$ has the same expression as in (18), and

$$\begin{aligned}
I_4(\lambda) &= \int_0^1 \delta^{e_1-1}(1-\delta)^{e_2-1}(1-\delta+\lambda\delta)^{-e_3} d\delta \\
e_1 &= a_{.} \\
e_2 &= b_{.} \\
e_3 &= a_1 + b_1
\end{aligned}$$

Following the expressions for $m(\lambda \mid x)$ and $m_A(\lambda \mid x)$ in (16), (17),(20) and (21), the evaluation of the loss of information measures requires firstly numerical integration on the rectangle $(0, 1) \times (0, \infty)$ and secondly Montecarlo simulation to compute the expected values. The numerical integrations are carried out by the gaussian quadrature method applied in a first step to the interval $(0, 1)$ and in a second step to the interval $(0, \infty)$. Because of the high cost of these double numerical integrations, we have considered a (n, n_0)-grid just fine enough for providing insights about the general behavior of the information loss in the (n, n_0)-space. All the simulation have been performed with programs written in C language which call subroutines in Fortran language for the pseudo-random number generation (Ranlib package, Brown and Lovato (1991)) and for the numerical integration (Quadpack routines, Piessens et al (1983)). The figures were made in Splus 3.1. Some illustrative results along with the main conclusions are presented in next section.

3 Results

In this section we present some details on results briefly sketched in Mouchart and Scheihing(1993) and relative to the comparison based on discrepancy measures; we also present new results regarding the measure of relative efficiency along with some comparison of the behavior of these two measures.

3.1 The discrepancy measures

For both the multinomial sampling and the independent binomial sampling, the evaluation of the Hellinger distance and the Entropy divergence gave roughly similar results for the range of sample sizes and precision of prior information we have simulated. Figure 1 presents, as an example, the case of the Hellinger distance for the multinomial sampling. Let us consider the loss of information $L(n, n_0)$, defined in (5), as a function of n, the sample size, and n_0, the precision of the posterior distribution. Figure 1(a) suggests that $L(n, n_0)$ behaves approximately as:

$$L(n, n_0) \sim h(\frac{n}{n_0}) \tag{22}$$

where $h : I\!R_+ \to I\!R_+$ is such that $h' > 0$ or equivalently: $L(n, \alpha(c)n_0) = c$ with $\alpha : I\!R_+ \to I\!R_+$ is such that $\alpha' > 0$. This suggests that $L_n^{'} > 0$ and $L_{n_0}^{'} < 0$, implying in particular that the loss of information does *not* vanish asymptotically (in n). The graphs also suggest that $L_{n,n}^{''} < 0$, i.e. for any n_0, $L(n, n_0)$ is concave in n (Figures 1(b)) and $L_{n_0,n_0}^{''} > 0$ i.e. for any n, $L(n, n_0)$ is convex in n_0 (Figures 1(c)) and eventually tends to 0 when $n_0 \to \infty$ (in which case the prior distribution becomes dogmatic and remains unrevised by the sample, whether the revision is "exact" or "approximate"). Thus, in this case, the error of approximation tends to vanish when $\frac{n}{n_0} \to 0$ or when $n_0 \to \infty$ for fixed n whereas this error does not tend to zero, but rather tends to increase when the sample size $n \to \infty$ with fixed prior precision n_0.

3.2 The relative efficiency

Figure 2 reports the results on the multinomial sampling. Similarly to the comparison in terms of discrepancy, the relative efficiency appears to be decreasing with the sample size n, i.e. the quality of the approximation deteriorates as n increases.

The effect of increasing the precision of the a priori information, n_0, is however more complex. The non monotone behavior shown by figure 2 may be understood if we look to the two expected values involved in the computation of the relative efficiency:

$$f_1(n_0) = E_x[(E(\lambda \mid x) - E^A(\lambda \mid x))^2] \tag{23}$$

$$f_2(n_0) = E_x(V(\lambda \mid x)) \tag{24}$$

In general terms, f_1 and f_2 are positive functions of n_0 decreasing to 0 when n_0 goes to infinity. On the other hand, figure 3 suggests that f_1 is smaller than f_2 and that they decrease at different speeds: for smaller values of n_0, the rate of decrease of f_2 is higher than f_1, and for greater values of n_0, the rate of decrease of f_2 is lower than f_1. Then the value of n_0 where both decreasing speeds have the same value correspond to the maximum for the ratio between f_1 an f_2 and therefore to the minimum for the relative efficiency. Finally, as in the discrepancy measure case, the results suggest that the error of approximation tends to vanish when $\frac{n}{n_0}$ goes to 0, that is the approximation can be qualified as a "small sample approximation".

3.3 Remarks on the simulation accuracy

Let us mention that the graphs relative to the Hellinger distance appear to display a more regular pattern than those relative to the entropy divergence, particularly for the case of independent binomial samplings. These irregularities are due to simulation problems. In general terms, one may suspect that the multinomial sampling generates more variations than the independent binomial sampling because this one may be obtained as a conditioning (on one margin) of the first one. One may also suspect more variation for the Entropy divergence because it is not compact-valued as is the Hellinger distance.

In the case of the relative efficiency, the graphs are based on simulation of size $N = 10000$ and $N = 5000$; previous trials with $N = 1000$ revealed dramatic irregularities.

Figure 4 may also help getting more precise an insight on the simulation problems. The coefficient of variation associated to the simulations is plotted as a function of a expected Hellinger distance(figure 4(a)) and as a function of the two expected values involved in the Relative Efficiency evaluation (figure 4(b) and (c)). Note that in the case of Hellinger distance there is a clear negative association between the coefficient of variation and the average (i.e. a smaller average is associated to a higher relative dispersion). On the other hand, in the case of the Relative Efficiency, a positive association between the coefficient of variation and the expected values is found. In particular, we chose higher simulation sizes for smaller values of n_0 (see table 1) because higher expected values were found in that case.

References

[1] Brown B.W. and J. Lovato (1991. Ranlib: Library of Fortran Routines for Random Number Generation. (bwb@odin.mda.uth.tmc.edu).

[2] Florens, J.-P., Mouchart, M., and J.-M. Rolin (1990). *Elements of Bayesian Statistics*, Marcel Dekker, inc., New York.

[3] Mouchart, M. and E. Scheihing (1993). Fisher test and inference in 2x2 contigency tables: A bayesian evaluation. *Proceedings of the section on Bayesian Statistical Science of American Statistical Association*, pp. 235-240.

[4] Piessens R., de Donker E. and D. Kahaner (1983). Quadpack: A package for the Numerical Computation of Definite Univariate Integrals. Appl. Math. and Prog. Div. - K.U. Leuven, Belgium.

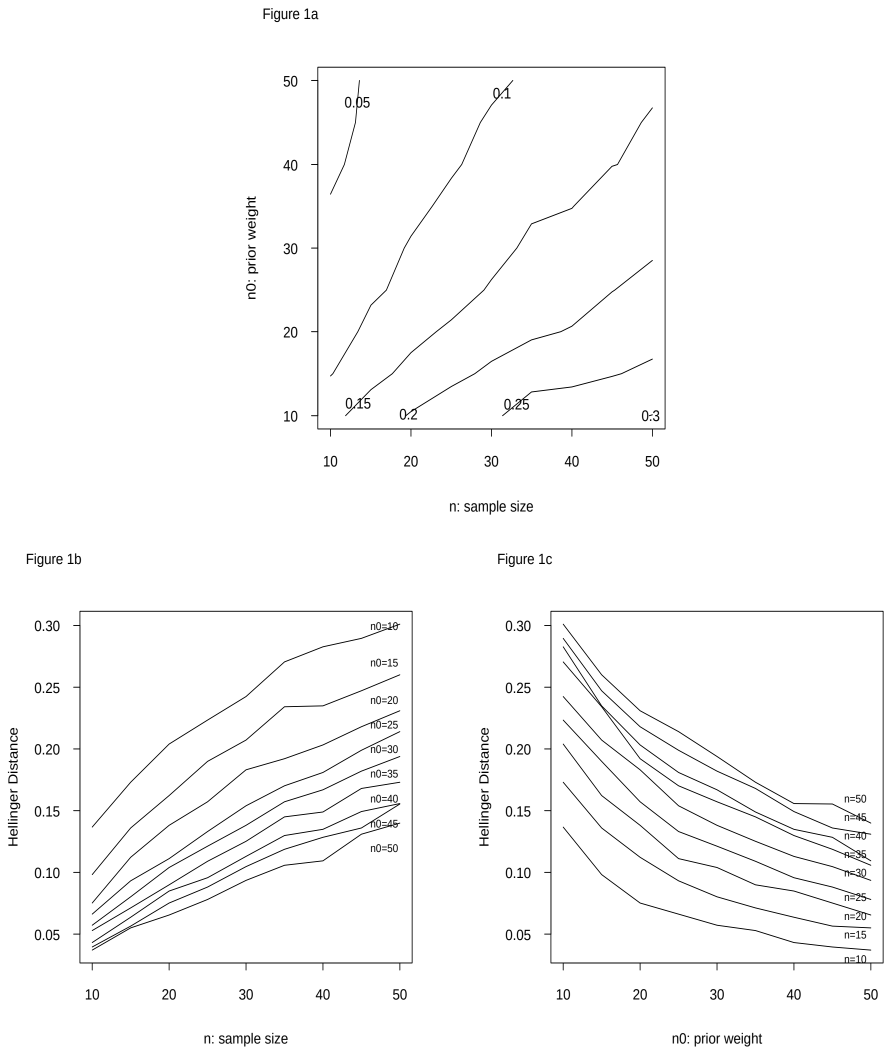
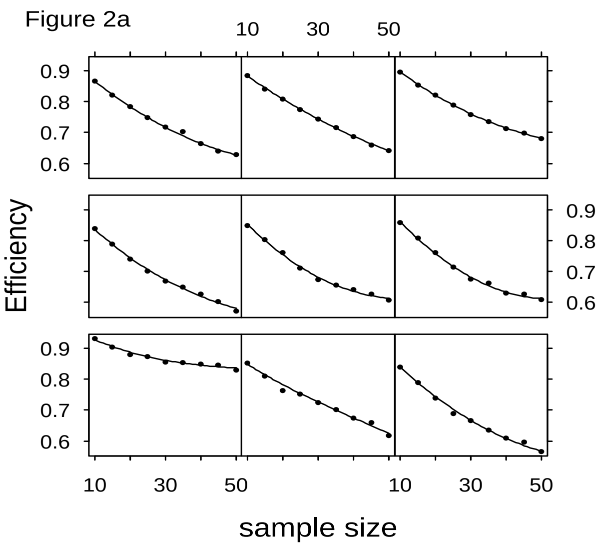


Figure 1: Expectation of the Hellinger distance in the multinomial sampling as a function of the size sample n and the prior weight n_0. Contour (fig. 1(a)) and one-dimensional plots (fig. 1(b) and (c)).



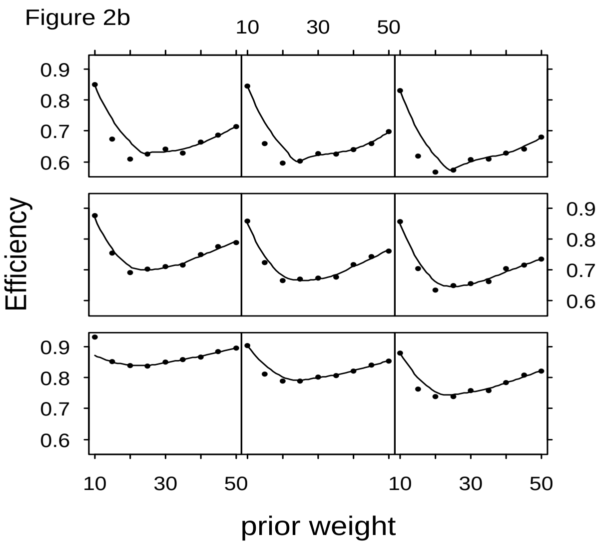
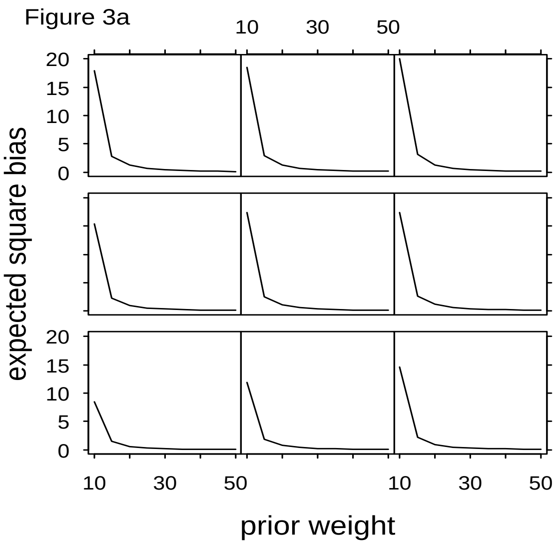


Figure 2: Relative Efficiency in the multinomial sampling as a function of the sample size for fixed values of the prior weight (fig. 2(a), from the left bottom plot corresponding to $n_0 = 10$ to the right top plot corresponding to $n_0 = 50$) and as a function of the prior weight for fixed values of the sample size (fig. 2(b) in analogous order).



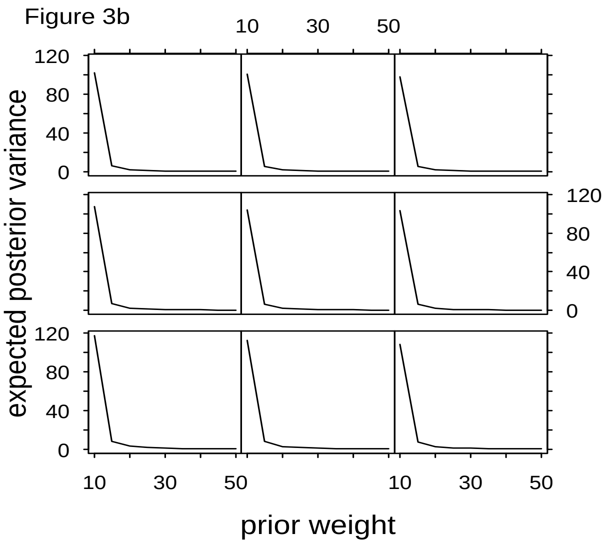


Figure 3: Expected values involved in the Relative Efficiency evaluation as a function of the prior weight n_0 for fixed values of the sample size n (from the left bottom plot corresponding to $n = 10$ to the right top plot corresponding to $n = 50$).

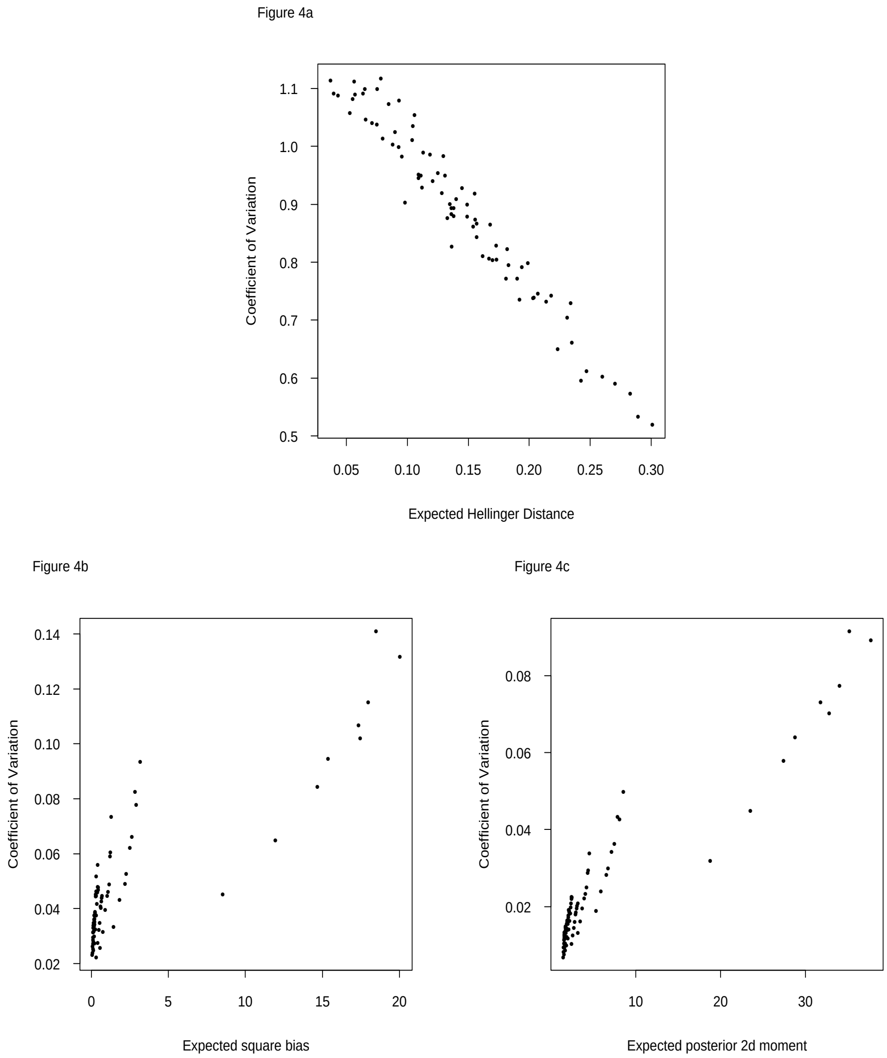


Figure 4: Coefficient of variation versus expected Hellinger distance (fig. 4(a)) and expected terms involved in the Relative Efficiency computation (figs. 4(b) and (c)), for the multinomial sampling.