Faculty Research

University of Michigan Business School

WORKING PAPER SERIES

On Priors with a Kullback-Leibler Property

Stephen Walker
University of Bath

Paul Damien
University of Michigan Business School

Working Paper 02-007

On Priors with a Kullback-Leibler Property

STEPHEN WALKER* AND PAUL DAMIEN**

*Department of Mathematical Sciences, University of Bath

**University of Michigan Business School

ABSTRACT. In this paper, we highlight properties of Bayesian models in which the prior puts positive mass on all Kullback–Leibler neighbourhoods of all densities. These properties are concerned with model choice via the Bayes factor, density estimation and the maximisation of expected utility for decision problems. The results suggest it is appropriate to label a prior with this Kullback–Leibler property as a true Bayesian model. In our illustrations we focus on the Bayes factor and show that whatever models are being compared, the [log(Bayes factor)]/[sample size] converges to a nonrandom number which has a nice interpretation.

KEYWORDS: Bayes factor, decision theory, exchangeability, expected utility rule, Kullback–Leibler divergence.

1. Introduction. Recent Bayesian nonparametric literature has focused on consistency properties of Bayesian procedures. See, for example, Wasserman (1998), Barron, Schervish and Wasserman (1999), and Walker (2002). Based on the results from these papers, we argue that it is practically realistic to define a true Bayesian model as one in which the prior puts positive mass on all Kullback–Leibler neighbourhoods of all densities. Our reasons for this position are the theme and point of the paper.

We consider solely the case when f_0 is a density function and $X^n = (X_1, \ldots, X_n)$ are available as a random sample from f_0, the first n observations of a possibly infinite sequence $X_1, X_2, \ldots$. Since f_0 is unknown, the Bayesian constructs a prior distribution on the relevant space of density functions, or distribution functions, reflecting available prior information about the location of f_0. Assuming all the densities under consideration are dominated by some σ-finite measure, which we will take to be the Lebesgue

measure, Bayes theorem and the data X^n combine to update the prior to the posterior.

There are compelling reasons why a Bayesian should use a prior distribution which puts positive mass on all Kullback–Leibler neighbourhoods of all densities; in particular, on all Kullback–Leibler neighbourhoods of f_0. Obviously, if f_0 is unknown a priori, to guarantee the prior puts positive mass on all Kullback–Leibler neighbourhoods of f_0, it is required to put positive mass on all Kullback–Leibler neighbourhoods of all densities. We shall refer to this as the Kullback–Leibler property for the prior Π. In order to achieve this, a nonparametric prior is required. For specific examples of priors with the Kullback–Leibler property, see the recent paper by Barron, Schervish and Wasserman (1999). We should point out that all Bayesian models, that is, $M = \{f(x;\theta),\ \pi(\theta)\}$, define a prior probability Π on the space of density functions. A random density function from Π is chosen by first choosing a θ from π and putting $f(\cdot) \equiv f(\cdot;\theta)$. Hence, for us, a Bayesian model is precisely the prior Π. A parametric model of finite dimensions will not satisfy the Kullback–Leibler property, unless f_0 is known to belong to a particular parametric family.

The following reasons suggest that Π should have the Kullback–Leibler property:

1 . Many practising statisticians would argue that parametric models are sufficient when combined with model checking and model comparison diagnostics. See, for example, Bernardo and Smith (1994). However, Draper (1999), in an insightful discussion of the paper by Walker et al. (1999), points out that allocating probability mass one to parametric subsets of densities should not be done lightly. The reason being that on switching models, when the original model under consideration is found to be deficient in some sense, exposes the statistician to the very real possibility of poor calibration. Therefore, there is a very practical reason for assigning mass one to the set of all densities; the data can offer no surprises.

2. It is shown that if Π does have the Kullback–Leibler property then the Bayes factor comparing this model with any other model will always eventually support (under mild regularity conditions) the prior with the Kullback–Leibler property. The precise result is stated in Section 2. The conclusion is that there is no motivation to put the prior Π

under the scrutiny of a Bayes factor, unless it is with another prior which also shares the Kullback–Leibler property.

3. Decisions made via the maximisation of expected utility are consistent when using a prior with the Kullback–Leibler property. This is proved in Section 3. That is, with a utility function and f_0, there is a well defined correct action, unknown just as f_0 is unknown. Decisions are consistent if the decision rule eventually sticks on this correct action.

4. For those interested in density estimation, there exists a Kullback–Leibler consistent sequence of predictive densities based on a prior with the Kullback–Leibler property. This is precisely stated in Section 4.

Before proceeding, we introduce the notation used throughout the paper. We let Π^n denote the posterior distribution given X^n. Then define $I_n = \int R_n(f)\,\Pi(\mathrm{d}f)$, $n \geq 1$, and $I_0 = 1$, where $R_n(f) = \prod_{i=1}^n f(X_i)/f_0(x_i)$. Define $f_n = \int f\,\Pi^n(\mathrm{d}f)$ to be the predictive density, and also define $D(f) = \int \log(f_0/f)\,f_0$ to be the Kullback–Leibler divergence between f_0 and f. In the following, a.s. will be with respect to the infinite product measure F_0^∞.

2. Bayes factors. Bayes factors are widely used in Bayesian model selection problems. See, for example, Bernardo and Smith (1994) for a review. To date, asymptotic studies of Bayes factors have only been formulated when one of the models is "correct". See, for example, Gelfand and Dey (1994). The Bayes factor for comparing model 1 with model 2 is given by

$$B_n = I_{1n}/I_{2n},$$

where $I_{jn} = \int R_n(f)\,\Pi_j(\mathrm{d}f)$, and the Bayesian models are fully characterised by Π_1 and Π_2. Recall that all Bayesian models induce prior distributions on the space of density functions.

Bayesian models, characterised by Π, will be associated with a $\delta \geq 0$. This δ is such that $\Pi\{f : D(f) < d\} > 0$ only for, and for all, $d \geq \delta$.

THEOREM 1. (WALKER, 2002). If

1. $\Pi_j\{f : D(f) < d\} > 0$ only for, and for all, $d > \delta_j$.

2. $\sum_n n^{-2}\mathrm{Var}(\log I_{jn}/I_{j\,n-1}) < \infty$

3. $\liminf_n D(f_{jn}) \geq \delta_j$ a.s.

then

$$n^{-1} \log B_n \to \delta_2 - \delta_1 \text{ a.s.}$$

Consequently, $B_n \to \infty$ a.s. (preferring model 1) if, and only if, $\delta_1 < \delta_2$. This makes sense. Note that the rate will be exponential; that is, $B_n \sim \exp\{n(\delta_2 - \delta_1)\}$. Obviously, if $\delta_1 = 0$, then $B_n \to \infty$ a.s. for all $\delta_2 > 0$. Condition 2. is an extremely mild condition to be satisfied. Condition 3. is also a realistic assumption to make; one would not anticipate the predictive density to get closer than δ to f_0 in a Kullback–Leibler sense, if the prior has no densities this close in the Kullback–Leibler support. We present illustrations of theorem 1 in Section 5.

If Π has the Kullback–Leibler property and the competing model does not, then the Bayes factor will eventually prefer the model which does have the Kullback–Leibler property. A model with this property can therefore rightly be defined as a true model. There is no point in comparing it with a model which does not have the Kullback–Leibler property.

3. Bayes decision theory. Here we provide further support for the notion that a prior Π with the Kullback–Leibler property can be called a true model.

Taking the notation from Hirshleifer and Riley (1992), the elements of a decision problem are as follows :

(1) a finite set of actions indexed by a and for practical purposes we assume $a \in \{1, \ldots, N\}$, for some integer N. While much of decision theory is written up with the notion of a continuous set of actions, in practice the number of decisions that can be made are finite, see Lindley (1985) for a discussion.

(2) a set of states of nature, which we take to be the appropriate space of distribution functions, say $\mathcal{G}$. [1]

(3) a consequence function $c(a, F)$ showing outcomes under all combinations of actions and states of nature.

[1] We assume that the relevant unknown state of nature is the distribution generating the data. This gives us a general framework to work with. Certainly, knowing the true distribution will solve all decision problems associated with the data.

(4) a preference scaling function $v(c)$ measuring the desirability of the consequence c.

(5) a probability distribution on $\mathcal{G}$ representing beliefs in the true state of nature. In a Bayesian context this probability is the prior Π in the no sample problem and is Π^n once the data X^n has been observed.

The Von Neumann–Morgenstern (1947) *expected–utility rule* then asserts that the best decision is to take the action a which maximises

$$U_n(a) = \int v\{c(a, F)\}\, \Pi^n(\mathrm{d}F).$$

This expected–utility rule is applicable if and only if the $v(\cdot)$ function has been determined in a particular way which leads to $v(c)$ being bounded, specifically $0 \leq v(c) \leq 1$. That is, the $v(c)$ has a probabilistic interpretation. See Hirshleifer and Riley (1992). There are differing opinions on the point of a bounded elementary utility function. See for example De Groot (1970) who relaxes the axioms of Von-Neumann-Morgenstern. We would point out that with unbounded $v(\cdot)$, it is not guaranteed that $U_n(a)$ even exists, and since this depends on f_0 which is unknown, the bounded $v(\cdot)$ makes most sense.

It is not our intention to discuss the expected–utility rule further. Our aim is to show that if Π has the Kullback–Leibler property then the decision rule eventually sticks to the action which maximises $U_0(a) = v\{c(a, F_0)\}$, which can be classified as the correct action, obviously unknown because F_0 is unknown.

THEOREM 2. If Π has the Kullback–Leibler property then $U_n(a) \to U_0(a)$ a.s. for all a.

PROOF. If Π has the Kullback–Leibler property then Π^n converges weakly to Π_0 a.s., where Π_0 is the probability measure with point mass one at F_0. See Schwartz (1965). The Portmanteau theorem (see, for example, Billingsley, 1968, Theorem 2.1), then gives the desired convergence result for $U_n(a)$, assuming that v is suitably smooth.

Clearly, if $U_n(a) \to U_0(a)$ a.s. for all a then the maximiser over $U_n(a)$, say a_n, will obviously eventually stick to a_0, which maximises $U_0(a)$.

4. Predictive density. Here we demonstrate that if Π has the Kullback–Leibler property then there exists a Kullback–Leibler consistent sequence of densities f^n. That is, $D(f^n) \to 0$ a.s.

THEOREM 3. (WALKER, 2002) Suppose that Π has the Kullback–Leibler property and

$$\sum_n n^{-2}\mathrm{Var}(\log I_n/I_{n-1}) < \infty.$$

If

$$f^N = \frac{1}{N}\sum_{n=1}^{N} f_n,$$

then $D(f^N) \to 0$ a.s.

Hence, for those who see density estimation as an important statistical procedure, f^n is an easily available Kullback–Leibler consistent sequence of densities. Nonparametric predictive densities are often hard to construct but are not hard to sample from. So, if it is possible to sample from f_n then it is obviously also possible to sample from f^n.

If Π has the Kullback–Leibler property then the condition

$$\sum_n n^{-2}\mathrm{Var}(\log I_n/I_{n-1}) < \infty$$

is an extremely mild constraint. If $I_n = \exp(-nt_n)$ then $t_n \to 0$ a.s. if Π has the Kullback–Leibler property. See Barron, Schervish and Wasserman (1999). It is therefore sufficient that

$$\sum_n \mathrm{E}\{(t_n - t_{n-1})^2\} < \infty.$$

If, for example, $t_n = O(n^{-s})$ for some $s > 0$ then $|t_n - t_{n-1}| = O(n^{-1-s})$ and the condition will be easily satisfied.

5. Illustrations. Here we present four examples illustrating theorem 1.

EXAMPLE 1. In the first example we take the true density function to be $f_0(x) = \exp(-x)$. We take model 1 to be $f_1(x;\theta) = \theta\exp(-x\theta)$ with prior $\pi_1(\theta) = \exp(-\theta)$ and take model 2 to be fixed at $f_2(x) = 0.5\exp(-0.5x)$. It is easy here to see that $\delta_1 = 0$ and $\delta_2 = \log 2 - 0.5 = 0.193$. It is calculated that

$$\int \prod_{i=1}^{n} f(x_i)\,\Pi_1(\mathrm{d}f) = \frac{n!}{(1+s_n)^{1+n}},$$

where $s_n = \sum_{i=1}^n x_i$, and

$$\int \prod_{i=1}^n f(x_i)\, \Pi_1(\mathrm{d}f) = (1/2)^n \exp(-s_n/2).$$

Following a simulation of data from f_0, figure 1 plotting $n^{-1} \log B_n$ for $n = 1 \ldots 3000$ is presented at the end of paper, where the convergence of $n^{-1} \log B_n$ to the correct value of 0.193 is evident.

EXAMPLE 2. In the second example we consider the case when both models are wrong, in the sense that neither prior has the Kullback–Leibler property. We now take $f_0(x) = x \exp(-x)$ and keep model 1 as in the first example. The second model is Weibull, $f_2(x) = x\theta \exp(-\theta x^2/2)$ with $\pi_2(\theta) = \exp(-\theta)$. Then $\delta_1 = 0.116$ and $\delta_2 = 0.099$ and so $\delta_2 - \delta_1 = -0.017$. Again, a simulation was performed of $n^{-1} \log B_n$ and figure 2 plotting the convergence to the correct value is also at the end of the paper. It should be noted in this case that the convergence is very slow and we took 1,000,000 samples. The figure shows every 350th value of $n^{-1} \log B_n$.

EXAMPLE 3. In this example we take a nonparametric prior, not infinite dimensional, but with a large number of parameters. With samples from $[0, 1]$, we have model 1 to be $f_1(x; \theta) = \theta x^{\theta - 1}$ with prior $\pi_1(\theta) = \exp(-\theta)$. We take model 2 as a histogram on $m = 1,000$ bins, with each bin of length $1/m$. The density function is $f_2(x) = m q_k \, \mathbf{1}((k-1)/m < x < k/m)$ and we take $(q_1 \ldots q_m)$ to have a Dirichlet prior with parameters all equal to 1. Then

$$\int \prod_{i=1}^n f(x_i)\, \Pi_1(\mathrm{d}f) = \frac{n! \prod_{i=1}^n x_i^{-1}}{(1 + t_n)^{1+n}},$$

where $t_n = -\sum_{i=1}^n \log x_i$, and

$$\int \prod_{i=1}^n f(x_i)\, \Pi_2(\mathrm{d}f) = \frac{m^n \Gamma(m)}{\Gamma(n+m)} \prod_{k=1}^m \Gamma(n_k + 1),$$

where $n_k = \sum_{i=1}^n \mathbf{1}((k-1)/m < x_i < k/m)$. If f_0 is uniform on $[0, 1]$, then both $\delta_1 = 0$ and $\delta_2 = 0$. As can be seen from the simulation of $n^{-1} \log B_n$ in figure 3 at the end of the paper, the Bayes factor always prefers the parametric model although asymptotically it prefers neither model, since $n^{-1} \log B_n \to 0$ a.s., although the convergence is extremely slow. This exposes the myth that Bayes factors always select the more complex model.

EXAMPLE 4. This is a slight variation of example 3. Here we retain model 2 and f_0 as in example 3 and take $f_1(x) = 2x$ to be fixed. Then $\delta_1 = 0.306$ and figure 4 shows the convergence of $n^{-1} \log B_n$ to -0.306. Again, very slowly. Note in this case that the Bayes factor always prefers the nonparametric model.

6. Discussion. In this paper, we have demonstrated how the Kullback–Leibler property for a prior Π provides good large sample properties for a number of Bayes procedures. We argue that Bayesians should be constructing priors with the Kullback–Leibler property, at the very least when there is doubt about the underlying shape of the density function generating the data. Although the results are based on large samples, the notion of having all densities in the Kullback–Leibler support of the prior must be an appealing one for all Bayesians. Indeed, from the Bayes factor perspective, there is no reason to compare a model with the Kullback–Leibler property with any other model, and so practically speaking meets the requirements of a true model. Barron, Schervish and Wasserman (1999) demonstrate that a number of nonparametric priors which are in use, such as Pólya trees and infinite dimensional exponential families, do have the Kullback–Leibler property.

For those interested in subjective issues, consider the following. Walker et al. (1999) show that it is possible to take subjective information from a parametric model and incorporate it into a nonparametric model. Then, for those who would acknowledge the existence of f_0, this paper demonstrates the practical relevance of a nonparametric model. For those who would not accept that an object such as f_0 does exist, the nonparametric approach using the Kullback–Leibler property offers the surprise-free approach (see point 1. in Section 1), and at a minimum avoids the poor calibration that may confront the statistician who is happy to hand out probability one to a host of possible models. See Draper (1999) for a detailed discussion of this point.

For those concerned with working in high dimensional spaces, the message from the collection of applied papers edited by Dey et al. (1998) is that it is no more difficult to routinely implement Bayesian nonparametric procedures than parametric ones, following the advent and rapid growth of user-friendly Markov chain Monte Carlo methods.

Other ideas for avoiding the model merry-go-round include Bayesian model averaging (Draper, 1995) and model selection, both ideas based on

a fixed set of models with associated probabilities of plausibility, rather than probabilities of correctness. Practically speaking it may not be difficult to assign probabilities to models; if there is a finite set then assigning equal probability is one option. A number of recent researchers have pointed out that model averaging usually outperforms model selection, and intuitively it is easy to see why this might be the case. We see model averaging as an attempt to construct a prior with large support (the idea being that at least one of the models may be close to f_0) using a collection of parametric models, and this could be seen as equivalent to a Bayesian nonparametric statistician who makes finite the infinite dimensional nonparametric model. This often happens, such as in the case of Pólya trees and the infinite dimensional exponential family; indeed it is necessary in these cases.

Acknowledgments. The work of S. Walker is financially supported by an EPSRC Advanced Research Fellowship.

References.

BARRON, A., SCHERVISH, M.J. and WASSERMAN, L. (1999). The consistency of distributions in nonparametric problems. *Ann. Statist.* **27**, 536–561.

BERNARDO, J.M. AND SMITH, A.F.M. (1994). *Bayesian Theory.* Wiley & Sons.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley & Sons.

DEY, D., SINHA, D. and MÜLLER, P. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics.* Lecture Notes in Statistics. Springer. NY.

DRAPER, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society Series B* **57**, 45–97.

DRAPER, D. (1999). Discussion of the paper "Bayesian nonparametric inference for random distributions and related functions" by Walker et al. *Journal of the Royal Statistical Society Series B* **61**, 485–527.

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

DE GROOT, M. (1970). *Optimal Statistical Decisions.* McGraw Hill Book Company.

HIRSHLEIFER, J. AND RILEY, J.G. (1992). *The Analysis of Uncertainty and Information.* Cambridge University Press.

LINDLEY, D.V. (1985). *Making Decisions* (2nd edn). Wiley & Sons.

LOÈVE, M. (1963). *Probability Theory.* 3rd Edn. D. Van Nostrand Company (Canada) Ltd.

SCHWARTZ, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete.* **4**, 10–26.

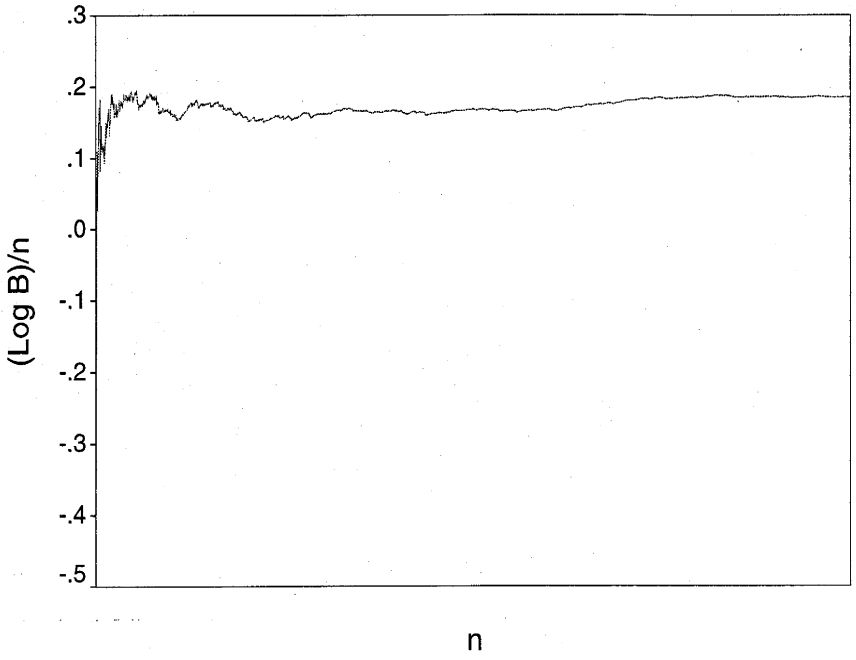
VON NEUMANN, J. and MORGENSTERN, O. (1947). *Theory of Games and Economic Behaviour* 2nd edn. Princeton University Press. Princeton N.J.

WALKER, S.G., DAMIEN P., LAUD, P.W. and SMITH, A.F.M. (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society Series B* **61**, 485–527.

WALKER, S.G. (2002). A new approach to Bayesian consistency. Submitted.

WASSERMAN, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. In *Practical Nonparametric and Semiparametric Bayesian Statistics* (D. Dey, P. Müller and D. Sinha, eds.), 293–304. *Lecture Notes in Statistics*, Springer. NY.

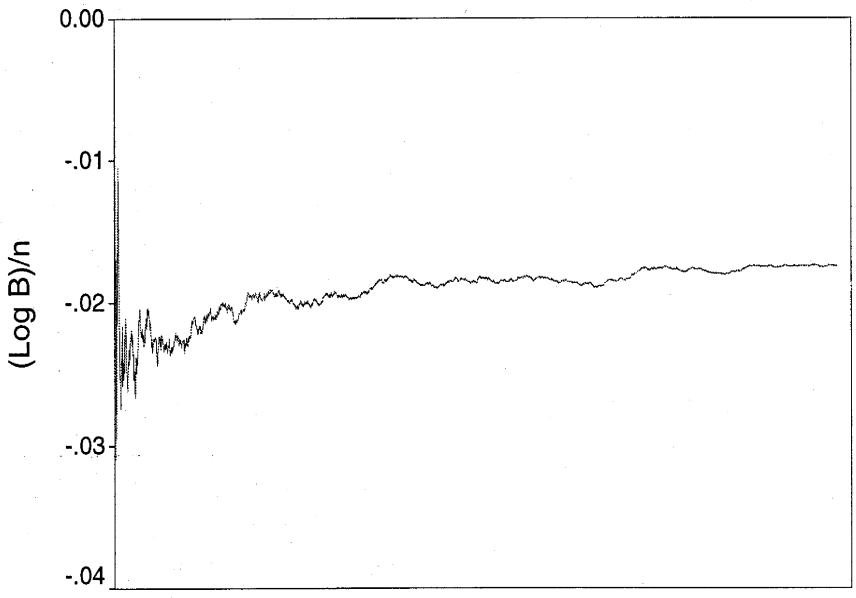
Example 1: Convergence of Bayes Factor



(Log B)/n --> 0.193

n = 1 to 3000

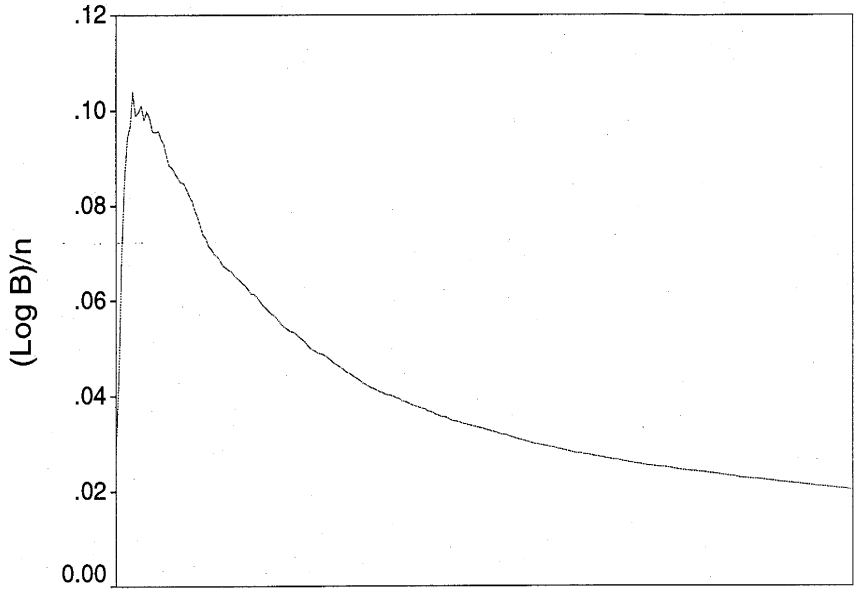
Example 2: Convergence of Bayes Factor



(Log B/)n --> -0.017

n = 1[350] 1000000

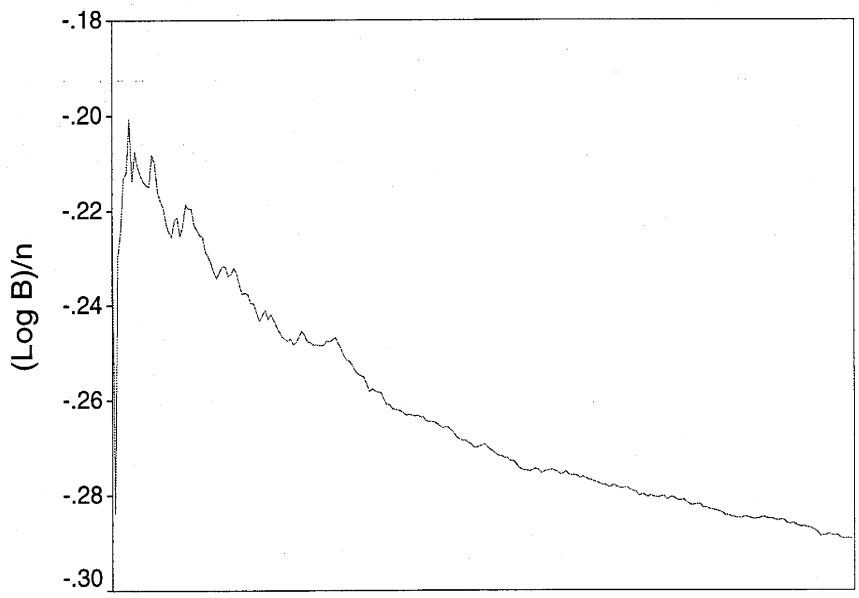
Example 3: Convergence of Bayes Factor



(Log B)/n --> 0

n = 1[350] 1000000

Example 4: Convergence of Bayes Factor



(Log B)/n --> -0.306

n = 1 [350] 1000000