AN EXAMINATION OF BAYESIAN METHODS AND INFERENCE: IN SEARCH OF THE TRUTH

by

Anirban DasGupta
Purdue University

Technical Report #94-16

Department of Statistics
Purdue University

June 1994

△

AN EXAMINATION OF BAYESIAN METHODS AND INFERENCE: IN SEARCH OF THE TRUTH*

Anirban DasGupta
Department of Statistics
Purdue University
West Lafayette, IN 47907-1399

Abstract

This is an article on common grounds between Bayesian and traditional or frequentist statistics and on behavioral evaluation of Bayes methods and procedures. Although many theorems are stated, we have deliberately omitted proofs. We have deliberately written this article in the style of telling a story. After a short expository section of six examples on common grounds, we discuss, in quite good detail four specific issues: sampling distributions of Bayes estimates in fixed samples, classical confidence of Bayesian intervals in fixed samples (including particular emphasis on how to match actual elicited information with a functional form of the prior), bias of Bayes estimates in fixed samples and its evaluation, estimation and correction including a general theory for evaluation for Gaussian data, various methods of estimation including the Bootstrap, many explicit examples, and the very important issue of whether it is too dangerous to attempt a bias correction, and finally the practically important issue of what should be a correct Bayesian formulation of the sample size problem, followed by actual sample sizes in the one way ANOVA problem, and a discussion of whether Bayesians should work out their sample sizes separately or just use widely available classical sample sizes. We finish with a brief section of concluding remarks. For easy location of topics, a table of contents is provided at the beginning. Relevant references are given only in the bibliography; this was necessary to meet a time constraint.

*This research was supported in part by NSF Grant DMS-93-077-27 at Purdue University.

TABLE OF CONTENTS

1. INTRODUCTION

The frequentist and Bayesian paradigms are different ways of doing statistics; from the point of view of foundations, they are fundamentally different, as is well known. In recent years, perhaps somewhat unfortunately advances in frequentist and Bayesian statistics have taken place to a significant extent in isolation of each other. It is our intention in this article to not go into differences between the two paradigms, but emphasize common grounds. It is indeed true that there are remarkable connections between how Bayesian and classical statistics are done.

It is impossible to go into the general issue of synthesis of classical and Bayesian statistics in detail. There is simply too much to be said, and too much for one person to know. After giving a brief introduction to the similarities and the connections between the two paradigms in section 2, we will go into consideration of three particular topics; one of these has been considered in the literature seriously, although more or less exclusively in the context of large samples. The other two have not been addressed seriously, although the general issues are well known. We will consider the following topics: classical coverage probability of Bayesian confidence (credible) intervals when the sample size is not necessarily large - in particular what kind of priors (informative priors) will give satisfactory classical coverage probabilities. The second topic we will discuss is the issue of bias of Bayes rules and subsequent bias correction. The third topic is the issue of sample sizes - practitioners quite routinely use sample size prescriptions in standard problems as guidelines in practical studies. These are all classical sample sizes, however; i.e., usually one specifies a desirable power and an acceptable type 1 error probability and a chart or a table gives a sample size consistent with these needs. The formulation in the Bayesian context would be different; now one needs a small posterior error probability. We will talk about two things: what happens if classical sample sizes are used, and what is the Bayesian sample size in the correct Bayesian formulation.

It is well known that Bayes rules generally are not unbiased. Barring that, no serious effort has been made to evaluate their bias, or to study the issue of whether some bias correction should be done, and if so how. It is clear that the three topics we discuss are extremely general in nature; one can ask these questions in any problem where these

quantities make sense (coverage probabilities make sense in set estimation problems; one talks of bias generally in point estimation problems, etc.). We will necessarily have to orient our discussion into particular problems, however; this is for the sake of providing concrete information. A very general discussion of these topics obviously will not lead to anything directly informative. We will describe the particular contexts in the relevant sections. The topic of coverage probability of Bayesian confidence intervals has a good amount of literature, although in the context of large samples. Our focus is on fixed samples.

2. COMMON GROUNDS AND CONNECTIONS: A BRIEF INTRODUCTION

This section is independent of the rest of the article. The purpose is to give a general discussion, mostly of expository nature. In the process, we will give some new examples on connections between the two paradigms.

Example 1. Testing a point null hypothesis. The classical procedure in this problem is a Neyman-Pearson α level test. In the Bayesian paradigm, one needs to specify a loss function, and rejects or accepts the null hypothesis according as whether the posterior expected loss of accepting is more or less than that of rejecting the null hypothesis. It turns out that formally such a procedure is itself a Neyman-Pearson test of a suitable level. This has been sometimes used to recommend that the choice of an appropriate α level should be done from considerations of the risk associated with false acceptance or rejection and prior odds for the two hypotheses.

Example 2. Maximum likelihood estimates. Maximum likelihood estimates are values of an unknown parameter that maximize the likelihood function. Verbally, one can think of the MLE therefore as the parameter value that best explains an observed phenomenon. Formally, one can think of an MLE as providing the model that minimizes the distance between a fitted model and the true model. In a Bayesian framework, one needs a prior distribution (density) on the unknown parameter before one can even start. Complete ignorance about the parameter is sometimes formulated by adopting a uniform prior on the unknown parameter. In such a case, it follows right away that the posterior density is proportional to the likelihood function and therefore the most likely value of

the parameter, i.e., the posterior mode, is exactly the MLE. More sophisticated connections between the MLE and Bayesian estimates can be given. The popular estimate in the Bayesian world is not really the posterior mode; it is the posterior mean. A more meaningful question may therefore be if the posterior mean for a uniform prior coincides with the MLE. In a rather technical article, the following result with apparently very promising computational implications was proved:

Theorem (DasGupta). Take n observations from a multivariate normal distribution with an unknown mean vector which we assign a uniform prior. We want to estimate a function g of the unknown mean. Then the MLE of g coincides with the posterior mean of g for all samples $X_1, \ldots, X_n$ and all sample sizes n if (and only if) g is a Harmonic function.

Since there are plenty of functions which are Harmonic in more than one dimension, this says that the classical ML approach and the uniform prior Bayes approach will give EXACTLY the same answer all the time for plenty of functions. The promise is, however, not only in direct application of this result, but also in the following general approximation it points to.

Take a function h that is not Harmonic. Also take a general prior that is not necessarily the uniform prior. Suppose the sample size n is not too small. Then the following sequence of approximations can be written (and made formal, from a mathematician's standard):

$$Eh(\underset{\sim}{\theta})/\underset{\sim}{x} --(\theta \ \underset{\sim}{)}/\underset{\sim}{\theta} \sim N(\underset{\sim}{x}, \frac{1}{n}I) \approx Eh(\underset{\sim}{\theta})I(\underset{\sim}{\theta} \in K)/\underset{\sim}{\theta} \sim N\left(\underset{\sim}{x}, \frac{1}{n}I\right)$$

$$\approx Eg(\underset{\sim}{\theta})I(\underset{\sim}{\theta} \in K)/\underset{\sim}{\theta} \sim N\left(\underset{\sim}{x}, \frac{1}{n}I\right) \approx g(\underset{\sim}{x}),$$

where K is an appropriate large compact set and g is the best harmonic approximant to h on the compact set K.

The construction of the Best Harmonic approximant of h on a sphere involves consideration of delicate mathematics, including the solution to Dirichlet problems on a sphere, but is well understood. It will be beside the point to wander into those issues here. The point is the MLE and the Bayes estimate are often approximately equal in parametric estimation problems. The actual result is quite a bit more general in that normally distributed data are not needed. Mixture models are included.

Example 3. The Fisher Discriminant Function. Suppose one has two multivariate normal populations $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$. Fisher proposed the linear function$(\mu_1 - \mu_2)'\Sigma^{-1}X$ as a criterion to discriminate between the two populations. This has now become a historically famous statistical method: widely used, perhaps even time tested. Fisher's original derivation was completely classical in nature. Certainly not even a hint of Bayesian thinking was present anywhere. The following connection therefore seems rather interesting.

Theorem (DasGupta, McCabe, and Mukhopadhyay). Consider a mixed population $(1-p)N(\mu_1, \Sigma) + pN(\mu_2, \Sigma)$. Suppose the mixing proportion p is not known. Assign p any prior $\pi(p)$. Then among all linear functions of the multivariate data vector X, the Fisher linear discriminant function maximizes the expected information (Fisher) about the mixing proportion p, where the expected information is with respect to the prior $\pi(p)$.

It is indeed curious how completely classical methods built from simple intuition or reasoning do have Bayesian connections.

Example 4. Nonparametric regression. In nonparametric regression, one assumes a model of the general form $y(t) = \mu(t) + \varepsilon(t)$, where $\mu(t)$ is in a broad class of regression functions and $\varepsilon(t)$ is an error. It has now become a standard practice to estimate the mean function $\mu(t)$ by using a combination of two criteria: good fitting, and reasonable smoothness. Smoothness is defined depending on the context. It is customary to have the function be (absolutely) continuous with a number of continuous derivatives: these are usually called Sobolev spaces.

A well known result is that the solution to such a constrained fitting problem is a smoothing spline; the order of the spline depends on how many continuous derivatives are assumed for the regression function.

The following Bayesian connection is also well known.

Theorem. Assume that the regression function $\mu(t)$ has the form of a polynomial plus a well defined error process $Z(t)$. The process $Z(t)$ is sufficiently smooth as a function of t. Suppose a fully uniform prior is given to the coefficients in this polynomial representation of the regression function $\mu(t)$. Then the pointwise posterior mean of the regression function is

a smoothing spline under standard conditions of zero correlation between various stochastic processes involved in the above modeling.

Again, the point is that the solution obtained originally from a purely algebraic point of view has a Bayesian interpretation. This connection has been used to use the posterior mse as an estimate of the sampling variance in this problem.

Example 5. The Sequential Probability Ratio Test. This is the sequential version of the classical likelihood ratio method for testing one point null hypothesis against another point null alternative. The motivation was that sometimes the full data are not required to choose between the two hypotheses; the evidence mounts soon in favor of one of them, and thus sequential sampling can save on cost.

The SPRT asks one to accept a hypothesis as soon as the likelihood ratio in its favor exceeds a given threshold level; otherwise the evidence is assessed to be insufficient and one continues sampling. The original paper of Wald gave pioneering calculations on the type 1 and type 2 error probabilities of this sequential test and on the savings on cost via calculation of average sample sizes at the decision stage.

The SPRT has the following Bayesian connection:

Theorem. Assign the two hypotheses a priori probabilities of π and $1 - \pi$ respectively, for $0 < \pi < 1$. Suppose the risks of false acceptance and rejection of the null hypothesis are 1 and K, for some positive constant K. Then the Bayesian test of this problem is a SPRT with appropriate threshold values determined by the constants π and K.

This, in spirit, is exactly the same as the corresponding fixed sample example we saw in Example 1. However, the proof is certainly harder.

We finish this section with the following example.

Example 6. Interpolation formulas of Numerical analysis. Many standard and famous methods of approximating a continuous function defined on a bounded interval,say [0,1], are based on some form of interpolation. Polynomial interpolation is naturally the first method to think of, but can be extremely inefficient in practice; even the choice of Chebyshev nodes is known to fail to provide uniform approximation in some cases, and other methods like the Bernstein operator have their theoretical value but are useless in

practice. Splines provide a family of functions which offer a lot of flexibility and efficiency at the same time. Interpolating cubic splines have now earned the status of perhaps the most common approximation method for implementation on computers. Suppose then that we have a function f in $C[0,1]$, which is "measurable"(in the everyday sense) at given points $x_1, \ldots, x_n$. We do not know f in a functional form (such problems arise quite routinely in consulting: f may be a response to a stimulant, deterministic but not known explicitly). The problem is to predict the value of f at a fixed point x on the basis of observations at n points $x_1, \ldots, x_n$. The following result is extremely interesting.

Theorem. Assume that f is a sample path of an once integrated Brownian motion. Suppose we treat this problem as a decision theory problem with mean squared error as criterion. Then the Bayes "estimate" of $f(x)$ given the data equals an interpolating spline of order 3 with knots at the data points.

This is only one result among many beautiful connections between classic numerical analysis tools and Bayesian statistics.

One can continue indefinitely; ridge regression, Kalman filters, almost any type of minimax estimation, etc. We proceed to the next section.

3. CLASSICAL CONFIDENCE OF BAYESIANLY CONSTRUCTED INTERVALS.

From a strict subjective Bayesian point of view, classical coverage probabilities of a Bayesian interval are not of relevance. One has a personal decision problem with a personal subjective prior and a personal utility function. The behavior of the corresponding Bayes procedure for observable (but unobserved) data is not important. It is simplistic, however, to conclude that the real world always acts or thinks that way. Most inferences are not meant for personal use only; they are used to make decisions as a group, to convince people and politicians, and the scientific soundness of a method is an inherently interesting question, which does relate to frequentist performance or typicality. These are not going to go away, certainly not in the immediate future.

Generally speaking a broad statement of the following type can be made and can be mathematically justified: if the subjective prior for the unknown parameter has a location

(like a strong prior guess), then Bayesian intervals corresponding to such a prior will tend to give excellent classical coverage probabilities if the prior guess is close or very close to being correct; however, at the same time, such classical coverage probabilities also tend to drop dramatically if the prior guess happens to be moderately incorrect. For people who value the assurance of the method to work in most cases, Bayesian intervals corresponding to very strong prior opinions thus come with a lack of insurance against ill conceived priors: if the strong prior opinion is correct, Bayesian intervals are great in every way; if they are incorrect, they have problems.

The following example merits mention.

Example 7. Estimation of a multivariate normal mean is one of the most common, most well studied, and may be even one of the most important problems of practical and theoretical statistics. In the simpler case of a known covariance matrix, the commonly employed estimate is the Hotelling ellipsoid centered at the sample mean vector (although simultaneous confidence intervals are more common in practice). In the Bayesian paradigm, there is a choice of many kinds of priors, but a multivariate normal conjugate prior is more common than anything else. The important thing is that a Bayesian with such a prior still uses an ellipsoid, but with two distinctions: the center is on the plane joining the sample mean and the prior mean, not at the sample mean exactly; also, the axes of the ellipsoid do not coincide with the axes of the contours of the normal distribution generating the data - the axes change due to the presence of another covariance matrix in the prior. In view of the preceding discussion, the following is not surprising:

Result. Denote the Bayesian ellipsoid by $S(X)$; then

$$P_\theta \{S(X) \text{ contains } \theta\} \to 0 \text{ as } || \underset{\sim}{\theta} || \to \infty,$$

and the convergence to zero is extremely fast, faster than the rate at which normal CDFs go to zero at the tails.

There is (justifiably) an impression among users that this problem can be eliminated by using noninformative priors. Noninformative priors in general tend to give Bayesian procedures that resemble classical procedures very closely or even exactly. There are two difficulties with such a resolution of this issue: virtually every result in this direction is of

relevance to large samples only, and furthermore the fundamental appeal of the Bayesian paradigm in its ability to use extraneous information outside of the data is totally lost in the adoption of objects like noninformative priors. There is also the very serious and very potent problem with interpretations: improper priors cannot be used to make uncertainty statements within the domain of standard probability theory, improper posteriors are an even bigger problem, and some of the most widely used simulation techniques like the Gibbs sampler can miserably fail in doing what we want it to do: give samples from a relevant probability distribution - the relevant distribution may not be a distribution because noninformative improper priors sometimes make the desired posterior just a measure, not a probability measure, yet the Gibbs sampler returns with samples. It is thus important on many grounds, philosophical and technical, to try to operate with informative or proper priors if possible. The following result is therefore very reassuring to people who value, for whatever reason, an assurance that a method works typically, irrespective of how the method was arrived at.

Theorem (Mukhopadhyay and DasGupta). For an unknown multivariate normal mean $\underset{\sim}{\theta}$, consider prior densities $g(\underset{\sim}{\theta})$ which satisfy the following (for now) technical condition:

$$|| \nabla g(\underset{\sim}{\theta}) || \ / \ g(\underset{\sim}{\theta}) \ \text{ is uniformly bounded.}$$

(There are lots of priors which satisfy this condition, and lot others which don't; normal priors do not satisfy this condition).

Take a scaled version of such a prior g:

$$g_c(\underset{\sim}{\theta}) = 1/c^p \cdot g(\underset{\sim}{\theta}/c),$$

where c is > 0, but will be usually somewhat large in applications.

Denote by $S_c(X)$ a $100(1-\alpha)\%$ Bayesian set estimator for $\underset{\sim}{\theta}$, under such a prior $g_c(\underset{\sim}{\theta})$; thus it is only the posterior probability of $S_c(X)$ which is $1-\alpha$; guarantee of satisfactory classical confidence needs to be dealt with as a separate issue.

Denote the classical coverage probability of $S_c(X)$ by $p(c, \underset{\sim}{\theta})$. Then,

$$p(c, \underset{\sim}{\theta}) = 1 - \alpha + 0(1/c), \ \text{uniformly in } \underset{\sim}{\theta}.$$

Discussion of the result: What does it mean? This result says that if one constructs a Bayesian set estimator for a multivariate normal mean with a prior of the type $g_c(\underset{\sim}{\theta})$, then a 95% (say) Bayesian set estimator comes with the guarantee that even its classical coverage probability is very very close to 95% AT EVERY $\underset{\sim}{\theta}$; unlike Bayesian ellipsoids which have extremely low classical confidence at values of $\underset{\sim}{\theta}$ where one or more of the coordinates is large, the Bayesian set estimator above has nearly 95% classical confidence whatever be $\underset{\sim}{\theta}$, provided the scaling constant c is somewhat large. There is no assumption of a large sample here.

However, the result needs to be understood very carefully; the following issues are obviously important:

a. Can we or should we choose c to be large just so it will give good classical confidence?

   The answer is 'no'. We should try to use this result in the following way: elicit the prior information that is there. Now match this information with a prior g of a functional form that satisfies the condition stated before. Finally , match the scaling constant c to be roughly consistent with the elicited information. Try the corresponding Bayesian set estimator for a few priors g of such kind and use one that seems to be the most suitable: in this consideration of which g is the most suitable, the classical confidence may also enter, if deemed an important factor for the users in that particular problem.

b. Does the same scaling constant c work whatever be the sample size?

   Again, the answer is 'no'. With a larger sample size, of course inference generally becomes more accurate. So a smaller sample will require a larger c to get the same kind of accuracy. However, as we said before, c should not be chosen large just because it is nice to do so. What this means is that even though the result is valid as a theorem whatever be the sample size, the strength with which the assurance of a 95% classical confidence is approximately correct does depend on the sample size n.

   The following concrete example may help.

Example 8. Bayesian intervals for a univariate normal mean. Suppose the prior information is that the median of $\underset{\sim}{\theta}$ is 0, and the quartiles are ± 2. Infinitely many priors are consistent with this limited information. They need not be symmetric at all. Depending

on the particular problem, one has to decide if a symmetric prior will be adopted because the quartiles are symmetric. Among symmetric priors, there are also infinitely many possible choices. The first inclination may be to use the $N(0, 8.78)$ prior; this matches the elicited information. But this is not a prior that satisfies the stated condition $|g'(\theta)|/g(\theta)$ is bounded. In fact, any normal prior does have a classical validity problem, as we saw before. So it is a good idea in this context to match the elicited information with another functional form. The Cauchy distribution centered at zero and with a scaling constant of $c = 2$ will satisfy the stated condition and match the elicited information. The corresponding Bayesian interval has to be found, however, at the expense of some numerical effort: there is no closed form formula if one adopts a Cauchy prior. But the numerical project is straightforward by the standards of computing today: repeated solution of an equation is necessary. The following table and the subsequent graph describes the classical confidence of our 95% Bayesian interval constructed from elicited information.

Table 1.

n	5	10	15	20	25	40
Minimum classical confidence	.9435	.9468	.9479	.9484	.9487	.9492

It is important to understand the very strong implication of these numbers. For a sample of size merely 5, a Cauchy prior matching the elicited information gives 95% Bayesian confidence and virtually 95% classical confidence at every θ as well, while a normal prior matching the elicited information would have given nearly zero classical confidence for values of θ about 2 standard deviations from the elicited median.

We have treated a very specific problem here. Important, but specific. It is the case that quite independent research will be necessary if the problem changes. In complex models with many parameters, the problem may well be too hard to progress. But the general moral of what we know and what we saw here in particular seems to be that matching elicited information with computationally convenient priors may lead to problems if classical confidences are relevant, for whatever reason. But that does not mean one has to throw one's hands up: there may be other informative priors which are computationally somewhat harder, but alleviate difficulties related to behavior as a method in repeated uses.

4. BIAS IN ESTIMATION BY BAYES RULES: EVALUATION, ESTIMATION, AND CORRECTION.

There are a number of reasons for interest in the bias of any estimate, including Bayes estimates. We will very soon list a number of reasons with quite concrete details. It is surprising, really, that no serious effort has been made to study this issue even in the most common or formally simple problems. There are probably two principal reasons for this:

a. Bayesians who use Bayes estimates more often than others are frequently not interested in bias because bias is not directly a Bayesian quantity;

b. Others who use Bayes estimates on other grounds probably realize how difficult the study of bias of Bayes estimates is as soon as one gets out of the domain of conjugate (or similar convenient) priors, no matter how computationally brilliant or ingenious one is. This is quite plainly a hard problem.

One can list many reasons for studying the bias of Bayes estimates; here are a few.

1. It is a reality that estimates which are off by a factor of 2 or 3 in the sense of an average are hard to sell; one can use all the argument about how the bias is being offset by incorporation of prior information, but it is simply not easy to convince users used to thinking of unbiasedness as a great virtue that bias by a large factor is not important.

2. Besides the issue of selling Bayesian methods, severe bias is actually bad from just common sense, if the Bayesian methods that are proposed are arising out of noninformative type automated priors. Automated priors come with practically the sole purpose of being used automatically, like a tool, a black box. A method that comes with the implicit recommendation of automated use, must be checked for typical behavior in automated use. Bias is one quantification of just that.

3. For a purely subjective (personal) prior in a personal problem, clearly the above argument for considering bias of the Bayes rule does not apply. However, even there, some formal reasons can be given for the importance of having a small bias. In the following, we give two results which indicate why this is so.

Theorem (DasGupta). A new general identity. Consider a generic estimation

problem with mean squared error as the criterion. Call the parameter to be estimated as θ, and the Bayes estimate under a given generic prior $\pi(\theta)$ as $d(X)$. Then the Bayes risk of $d(X)$ (i.e., the average MSE of the Bayes estimate) satisfies

$$r(\pi) = -\int \theta \cdot b(\theta)\pi(\theta)d\theta,$$

where $b(\theta)$ denotes the bias of the Bayes estimate $d(X)$.

Typically, in many applications, Bayesian estimates tend to be shrinkage estimates; thus the bias $b(\theta)$ and the parameter θ often have the same sign. What follows is that a bias large in magnitude does not cancel, but makes the Bayes risk large. Since a large Bayes risk is bad (or at least should be bad) from the Bayes viewpoint, this suggests that a large bias may be a red flag for even the Bayesian.

Here is another result in the same spirit.

Bayesian or not Bayesian, a statistician ought to be happy if his/her proposed estimate has the property that for practically all data, the estimate is virtually equal to the quantity being estimated. This is a statement of accuracy irrespective of data. One can think of this in the following manner: the statistician has the ideal goal of using an estimate $d(X)$ whose distribution is just a point mass at the parameter being estimated. Of course, this is unattainable; it is an ideal goal.It will therefore not be meaningless to consider the deviation of the distribution of an estimate $d(X)$ from this ideal goal.

Here the Bayesian runs immediately into an extremely hard problem. The distribution of a Bayes estimate, again with the exception of very convenient priors, is simply not something one can write on a piece of paper (the same can be said of many nonBayesian estimates as well). In some problems, and we emphasize that this is very specific on the problem, one can do tricky things. In principle, one can use numerical methods to compute the distribution, but this requires a nearly impossible amount of computing, and it is a very hard thing to determine if all the numerical errors, however small, made at various stages of this process keep the computed distribution reliable. And one needs to keep in mind that there is actually a family of distributions, one for each value of the parameter. We believe that for small values of the sample size n, there will be some asymmetry in the distribution of Bayes estimates in general, and various possibilities exist for accurately

approximating them. Saddlepoint approximations and Edgeworth expansions should also be looked at. In any event, the sampling distribution of the Bayes estimate is an interesting object, it needs much more investigation in small samples than what has been done, and the exact evaluation is going to be a hard problem (and particularly so in complex models).

Instead, let us try as a first step, an approximation. This is, as far as we see in some simple problems, quite good, but we have not tried any complex models at all. We approximate the distribution of the Bayes estimate $d(X)$ by a normal distribution with mean and variance equal to the mean and variance of $d(X)$, which are abstract quantities right now (t distributions may give still better approximations). The ideal goal is a distribution which is a point mass at the parameter. Common criteria for measuring deviation of one distribution from another do not work or work well with point masses - one gets uninteresting answers or no answers. We therefore also approximate the ideal goal; in general one can use what mathematicians call an approximate identity. In our case, mostly because we want to ultimately arrive at an answer that makes a point, we will use the normal distribution with mean at the parameter, and variance $= \varepsilon$, where ε is a small positive number.

Then, the following hold:

Result. The symmetrized Kullback - Leibler distance between the two normal distributions described above equals

$$\text{constant} \cdot \int (b^2(\theta) + v(\theta))\pi(\theta) + \text{ constant} \cdot \int v^{-1}(\theta)\pi(\theta)d\theta$$
$$+ \text{ constant} \cdot \int \frac{b^2(\theta)}{v(\theta)}\pi(\theta)d\theta - 1/2;$$

the constants refer to quantities free of the prior under consideration.

Thus a large bias, or a large bias in comparison to the variance at a value of the parameter important according to the prior keeps the ideal goal far from being achieved.

The following says the same thing with a different distance.

Result. The Hellinger distance between the two normal distributions described above satisfies

$$\int log(2 - d_H(\theta))\pi(\theta)d\theta \approx \text{constant} - 1/4 \cdot \int [\ logv(\theta) + b^2(\theta)/v(\theta)\ ]\ \pi(\theta)d\theta$$

Again, a bias large in comparison to the variance is seen to be bad in this formulation.

We will therefore now concentrate on the following issues:

a. Evaluation of the bias at a given value of the parameter;

b. Estimation of the bias (bias itself is a function of the parameter, and so needs to be estimated)

c. Correcting the Bayes rule for bias, including the issues of how and if.

We will first talk briefly about correction, giving a neat result, and return to the correction issue later in the section.

Theorem (DasGupta and Shyamalkumar). Consider a generic estimation problem with mean squared error as criterion and a given generic prior π. Call the parameter being estimated θ and assume $\int \theta^2 \pi(\theta) d\theta < \infty$. Denote the Bayes estimate of θ by $d_0(X)$ and its bias by $b(\theta)$.

Define the once corrected estimate $d_1(X)$ as

$$d_1(X) = d_0(X) - \hat{b}(\theta),$$

where $\hat{b}$ is the Bayes estimate of $b(\theta)$.

Inductively define the nth corrected estimate $d_n(X)$ as

$$d_n(X) = d_{n-1}(X) - \hat{b}_{n-1}(\theta),$$

where $\hat{b}_{n-1}$ denotes the Bayes estimate of $b_{n-1}(\theta)$.

Then the bias of the repeatedly corrected estimate $d_n(X)$ converges to zero (in L_2 with respect to the prior under consideration) if and only if the parameter θ being estimated is in the closed convex hull (in L_2 again) of the class of all parametric functions which are unbiasedly estimable.

Interpreted in a user's language, this means that repeatedly correcting an original Bayes estimate by always estimating the bias by a Bayes estimate will eventually make the bias zero if (and only if) the parameter has an approximately unbiased estimate, in which case the bias corrected Bayes estimate after a large number of corrections becomes approximately equal to an unbiased estimate.

The if and only if nature of this result suggests that there will be examples of quite natural estimation problems in which no amount of bias correction will eliminate the bias of a Bayes rule. The following is a simple example. After the example, we will give a strong generalization.

Example 9. Consider estimating the variance of a Bernoulli distribution with mean squared error as criterion and the noninformative uniform density as the prior for the unknown proportion p. Then the bias of the repeatedly bias corrected estimate always equals $1/6 - p(1-p)$, regardless of the number of bias corrections.

This generalizes to the following:

Theorem (DasGupta and Shyamalkumar). In the same problem, consider a general Binomial distribution, $\text{Bin}(n,p)$, and consider a general prior density $\pi(p)$. Let $\{P_k(p)\}$ denote the sequence of orthogonal polynomials with respect to the prior density $\pi(p)$. Let $g(p)$ be any parametric function which is orthogonal to the first n polynomials $P_1, P_2, \ldots, P_n$; then the bias of the repeatedly bias corrected Bayes estimate of $g(p)$ always equals a fixed function of p, independent of the number of bias corrections.

The bias itself, as a function of the parameter, is arguably the most interesting quantity in this context: without knowing the function or at least an idea of the function, it is hard to decide if we have the problem of a large bias at important parameter values and it is hard to parametrically estimate it (although other tools such as the Jackknife or the Bootstrap can be used and we will have an occasion to use these later in this section). As we mentioned earlier, exact evaluation of the bias as an explicit function of the parameter is an impossible problem, barring a few exceptional circumstances. As a starting step, it is a good idea to try to understand the severity of the bias of Bayes estimates in problems that are relatively more amenable to calculations that reduce the otherwise heavy computational burden. Let us try to get an understanding of this issue when data are coming from observations made at discrete times on a Gaussian process; the stationary case will be treated first; much later, we will show some results and calculations on Bayesian estimation of the drift of a Brownian motion. This is the only nonstationary case we will see in this particular article.

Suppose then that $X_1, \ldots, X_n$ are observations on a stationary Gaussian process with a mean μ and covariance function $\gamma(t)$ and we assume for the sake of nicety as well as a

latter Theorem that the process admits a spectral density $f(\lambda), -\pi \leq \lambda \leq \pi$. As one can expect, some exact results are possible with a Gaussian prior on the mean. Some close to exact results are possible with other priors as well; one such prior is the Double exponential which we will study later.

Once one starts to think about this problem, it becomes clear that the correct formulation is not quite a black and white issue. If we want to make statements of the kind "we are off by a factor of 2, or we are correct to within a factor of about 1.1", then clearly what is of concern is not the bias itself, but the relative bias $|b(\mu)|/|\mu|$. At this stage, the issue gets interestingly muddled. Conjugate (or similar) priors tend to keep the relative bias bounded (although it may still be large) but provide no control on the bias. On the other hand, priors one likes to think of as flat priors (safe priors??) keep the bias small but provide no control on the relative bias. The problem occurs at an unexpected place, namely when the mean is nearly zero. There is no resolution of this; may be flat priors are still better, with the understanding that if two numbers are both small, we will ignore the question of which is still smaller. Then relative bias stands on firm ground. The following holds.

Theorem (DasGupta). Denote the covariance matrix of the first n observations by Σ. Assume the mean μ has a Gaussian prior with mean 0 and variance τ^2. Then the relative bias of the Bayes estimate of μ is a constant independent of the parameter, and equals

$$\left|\frac{b(\mu)}{\mu}\right| = \alpha c, \text{ where } \alpha = \underset{\sim}{1}'\Sigma^{-1}\underset{\sim}{1} \text{ and } c = \frac{\tau^2}{\alpha\tau^2+1}.$$

Furthermore, assuming that the underlying process is not fully deterministic, i.e., $\int log f(\lambda)$ $d\lambda > -\infty, n \cdot \left|\frac{\tau^2}{\alpha\tau^2+1}\right|$ converges to $2\pi f(0)/\tau^2$, as $n \to \infty$.

Two comments are in order; notice that we have taken the mean of the prior to be zero. Conceptually, one obviously need not. However, if the prior has a nonzero mean, the relative bias is no longer bounded, even for finite n. Secondly, the result says that the relative bias is more serious if the spectral density is large at zero. This is also not surprising; it is a very well known fact that a large value of $f(0)$ is the cause of much trouble in most of these problems.

Example 10. AR(1) processes. The theorem says that in this case, for a large sample, the relative bias is approximately $1/n \cdot \sigma^2/\{\tau^2(1-\phi)^2\}$.

The danger of a very sharp prior manifests immediately; if $\frac{\sigma^2}{\tau^2}$ is taken to be 1, then this says that to achieve a relative bias of at most 5%, 222 observations will be necessary if the autoregression coefficient is a modest .7. The case gets increasingly worse as one approaches the unit root case, as expected.

Example 11. AR(2) processes. In this case, the coefficients must satisfy $\phi_1 + \phi_2 < 1$ for stationarity (this is only necessary). If $\phi_1 = .1$, and $\phi_2 = .7$, then the theorem says that 500 observations are necessary to keep the bias within a factor of 5%, if $\frac{\sigma^2}{\tau^2} = 1$ is used.

Example 12. Gaussian covariances. A popular model for a rapidly decreasing covariance function is $\delta(n) = \sigma^2 \cdot \exp(-\alpha n^2)$. This has been used in many applied areas, including spatial designs. The spectral density admits the representation

$$f(\lambda) = \frac{\sigma^2}{2\pi}\left[e^{\frac{-\lambda^2}{2\alpha}}\left\{\sum_{n=1}^{\infty} e^{-\alpha(n-z)^2} + \sum_{n=1}^{\infty} e^{-\alpha(n+z)^2}\right\} + 1\right]$$

where $z = \frac{i\lambda}{2\alpha}$.

The relative bias goes to zero at the exact rate $1/n \cdot \frac{\sigma^2}{\tau^2}\theta(\alpha)$ where $\theta(\cdot)$ is the Jacobi theta function evaluated at α. Again, as a specific illustration, $n = 50$ observations are necessary for a relative bias of 5% if $\frac{\sigma^2}{\tau^2}$ is taken as 1 and α is taken as .5108256238 (giving a lag 1 correlation of .6). All one needs is a table of the theta function.

Example 13. Equicorrelated observations. This is an interestingly exceptional case. The relative bias does not disappear with a large sample. In fact, the relative bias converges to $\sigma^2 \cdot \rho/\tau^2$, as the sample size $n \to \infty$. Thus a large correlation that refuses to taper off at large lags causes a large relative bias for the Bayes estimate, particularly if a sharp prior is used.

Example 14. A nonstationary example: Brownian motion with a drift. Consider the process $X_t = t\mu + \sigma B_t$, where μ is an unknown drift parameter, and B is standard Brownian motion starting at 0. We are interested in estimating the drift μ, which we assign a Gaussian prior with mean 0 and variance τ^2. Sometimes the drift parameter may be known to be positive, in which case the assumption of a Gaussian prior naturally will have to be changed. Although the assumption of stationarity is no longer valid, all the Gaussian structure is still in place, leading in a completely straightforward way to the following:

Result. Denote the vector $(1, 2, .., n)'$ by $\underset{\sim}{\alpha}$, and the matrix

$$\begin{bmatrix} 1 & 1 & \dots & \dots & & 1 \\ 1 & 2 & 2 & \dots & & 2 \\ 1 & 2 & 3 & 3 & \dots & 3 \\ . & . & . & . & & . \\ 1 & 2 & \dots & \dots & & n \end{bmatrix}$$

by Σ. Then the Bayes estimate of λ on the basis of n observations $X_1, \dots, X_n$ equals

$$d(X) = \underset{\sim}{\alpha}'\Sigma^{-1}\underset{\sim}{X}/(\underset{\sim}{\alpha}'\Sigma^{-1}\underset{\sim}{\alpha} + \frac{1}{\tau^2})$$

with a relative bias of

$$|b(\mu)|/|\mu| = 1/(\tau^2\underset{\sim}{\alpha}'\Sigma^{-1}\underset{\sim}{\alpha} + 1).$$

Together with the following interesting Lemma:

Lemma. $\underset{\sim}{\alpha}'\Sigma^{-1}\underset{\sim}{\alpha} = n$.

One has,

Result. For all $n, |b(\mu)|/|\mu| = 1/(n\tau^2 + 1)$.

A General Method for Bias Evaluation in Gaussian Processes.

Having earlier said that evaluation of the bias of a Bayes estimate as a function of the parameter is generally a hard problem, we are now going to show a new general theory for bias evaluation applicable to Gaussian processes. The following things should be made clear right at the start:

a. The theory is one for evaluating expectations of smooth functions of a multivariate random vector with a Gaussian distribution; it is more encompassing than bias evaluation of Bayes rules, therefore.

b. The theory gives an expansion for the expectation of a smooth function; successive terms have powers $n, n^2, \dots$ in the denominator when the expansion is applied to one dimensional data and n is the sample size. Typically, therefore, a few and usually two terms give an accurate approximation to the expectation being sought.

c. The derivation of the expansion is in an article of the author, DasGupta(1994); it is quite technical and uses facts from partial differential equations. We need not be concerned with it in this article.

d. Although the actual result applies to any number of dimensions and any covariance structure, here we will illustrate only a one dimensional problem, and therefore will only state the one dimensional result.

Theorem(DasGupta(1994)). Let $g: \quad R \to R$ be an infinitely differentiable function such that for all $k, m \geq 1, |g^{(k)}(x)| \cdot exp(-x^2/2) \cdot |x|^m \to 0$ as $|x| \to \infty$.

Let X be distributed as $N(\theta, t)$. Then,

$$E(g(X)) = g(\theta) + \sum_{k=1}^{\infty} (t/2)^k \cdot g^{(2k)}(\theta)/k!.$$

Example 15. As an illustration of the utility of this in our context, let us consider the following nontrivial Bayes problem. We have n iid observations from $N(\theta, \sigma^2)$ and the mean θ is given a Double exponential prior with density $exp(-|\theta|)/2$. Other scales for the prior can obviously be handled as well. The bias of the Bayes estimate $d(\overline{X})$ of the mean θ is by definition $E\{d(\overline{X})\} - \theta = E\{d(\overline{X}) - \overline{X}\}$.

At this stage a very nice and powerful identity due to Brown becomes useful. The Brown identity says the following:

Theorem. For a general prior π on the mean θ, the Bayes estimate of θ has the representation

$$d(\overline{X}) = \overline{X} + \sigma^2/n \cdot m'(\overline{X})/m(\overline{X}),$$

where $m(\overline{X})$ denotes the marginal density of $\overline{X}$, i.e.,

$$m(\overline{X}) = \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \int e^{\frac{-n}{2\sigma^2}(\overline{X}-\theta)^2} \pi(\theta) d\theta.$$

Thus, the bias equals $\sigma^2/n \cdot Em'(\overline{X})/m(\overline{X})$.

Identifying the function $m'(\overline{X})/m(\overline{X})$ with g in the Expansion stated above, and using $t = \sigma^2/n$, one has right away an expansion for the bias of the Bayes estimate. Carried to its full length, the expansion will give the exact bias. We want to terminate it at a small number of terms; in practice, at any given θ, one should stop when it becomes clear that incorporation of more terms is useless. Usually, just two terms suffice for any moderate value of the sample size n, like 10.

There is evidently a serious computational issue still to be addressed. The marginal density $m(\overline{X})$ needs to be differentiated five times for a two term approximation at every value θ of the parameter where the bias is to be evaluated. The truth is that for general priors, even the marginal density $m(\overline{X})$ itself can only be calculated numerically. The subsequent derivatives can therefore only be calculated by numerical differentiation, usually an unreliable computing exercise. In general therefore, it is better to evaluate the bias by simulation from the underlying normal distribution and repeated evaluation of $d(\overline{X})$ for each set of simulated observations. Sometimes, however, we get lucky. Certainly with normal priors, everything falls into place. However, for normal priors the entire problem is straightforward and there is no need to approximate in the manner suggested here. There are some other priors where again the computing burden is very significantly reduced due to the fortunate presence of exact formulas for the marginal density. The Double exponential prior we started with is one such case. An expression for $m(\overline{X})$ with such a prior is completely easy and well known. Repeated derivatives of $m'(\overline{X})/m(\overline{X})$ are just derivatives of $\log m(\overline{X})$, and these are also known exactly. Using these and a two term expansion as suggested above leads to an accurate approximation of the bias of the Bayes estimate, of which we give a plot with $\sigma = 1$ and $n = 10$. Although we have given only one example here, it is clear that this method will work (sometimes two terms may not suffice) in good generality. The only restriction is that one has to have the multivariate normal structure.

The plot of the bias of the Double exponential prior Bayes rule is instructive. Unlike conjugate prior Bayes estimates which have a linearly diverging bias, here the bias remains bounded and quite quickly attains an asymptote. The ease with which the theoretical expansion of the bias gives the function very accurately says that this method should be preferred whenever applicable.

Estimation and correction of the bias. Since simple closed form formulas are available for our use with conjugate priors, both problems are relatively easily tackled in that case. There is some question about the correct method to correct for bias, if correction is done. A Bayes estimate is more natural for the Bayesian to estimate the bias; with simple priors, this is fine. With other priors, the Bayes estimate of the bias will need numerical integration. This is not a very serious concern, if bias correction is done only once. To

do a second round of bias correction if necessary, it will be necessary to have some idea of how much correction was done the first time. This amounts to finding the bias of the once corrected estimate, and one way or the other calls for repeated numerical integration and evaluation of the Bayes estimate at many values: bias is an average quantity - replications are needed to find an average. An alternative to the Bayes estimate of the bias is an unbiased estimate of the bias. This is evidently immediately available without another round of computing.

Example 16. We investigated the potentials of bootstrap bias estimates in one example. $n = 15$ observations were simulated from an AR(1) process with mean μ and an autoregression coefficient of $\phi = .7$. We used again a Double Exponential prior on the mean, taking the quartiles to be 0 and 2. An exact analytical form of the Bayes estimate is available, and this was used. No numerical computing was therefore necessary to evaluate the estimate. For each given μ, the Bootstrap estimate of the bias was evaluated by using $B = 250$ Bootstrap replications. The exact bias as a function of the mean and the Bootstrap estimates of the bias are overlaid on a plot. We cannot say with confidence that this plot indicates strongly that Bootstrap estimates lack generally in accuracy in this problem. However, the plot indicates there may be a problem with both the bias and the accuracy of the Bootstrap bias estimates of Bayes rules. This is a very important issue and we believe it should be pursued to get a good understanding of the picture. Again notice that the general shape of the exact bias as a function of the mean is the same as for iid data: it is bounded, with an asymptote.

Associated with the issue of bias correction is another extremely serious question: should bias correction be done at all? The danger in a bias correction is that while it reduces the bias, it can cause moderate to major havoc to accuracy. In the present context, the appropriate measure of accuracy seems to be Bayes risk under the assumed prior. One needs an assurance that the increase in Bayes risk is inconsequential in comparison to the improvement in bias. With some straightforward calculations, one can quickly arrive at the following:

Result. Let α and c be as before. Suppose the bias of the original Bayes estimate of the mean μ is estimated by a Bayes estimate again. Then the once corrected esti-

mate equals $(2c - c^2\alpha)\ \underset{\sim}{1}'\Sigma^{-1}\underset{\sim}{X}$ and its bias for estimating the mean μ now improves to $\left((2c - c^2\alpha)\alpha - 1\right)\mu$. The Bayes risk of the once corrected estimate equals

$$(2c - c^2\alpha)^2\left[\frac{\alpha^2 c}{1 - cn} + \alpha\right] - (2c - c^2\alpha)\cdot\frac{2c\alpha}{1 - cn} + \frac{c}{1 - cn},$$

while the bias and the Bayes risk of the original Bayes estimate of the mean μ equal $(c\alpha - 1)\mu$ and c respectively.

This result will help decide if bias correction was wise. For instance, direct application of the collection of formulas above gives that if $\sigma = 1, \tau = 3, n = 35$, and one has equicorrelated observations with $p = .25$, then a single bias correction improves the bias by 97.07% and increases the Bayes risk by 2.84%. Usually the picture will not be so clearly in favor of a bias correction, but if it were so, it would seem wise to do it.

5. SAMPLE SIZE DETERMINATION: THE BAYESIAN PERSPECTIVE.

The problem of determining sample sizes that give an assurance of a prespecified accuracy (in whatever relevant way accuracy is defined) has assumed the status of textbook material in classical statistics. All among us and also people who use statistics have surely seen and used tables and charts of sample sizes that guarantee a power of .95 at a type 1 error level of .05 in standard hypothesis testing problems. We routinely teach our students in most elementary classes about such classical sample sizes. Several books and monographs testify to the value that practitioners assign to sample sizes as a preexperimental design component.

It is somewhat curious, therefore, that barring a recent outgrowth of interest and activity, determination of sample sizes correct according to a Bayesian formulation has been more or less a nonexistent topic in Bayesian research. First, one needs to understand that a new and careful formulation consistent with the Bayesian view of the world is necessary. The Bayesian has little primary use for sample sizes that assure low type 1 and 2 error probabilities. In the Bayesian's mind, the correct accuracy measure is a posterior accuracy: thus in a testing problem, a Bayesian would be naturally interested in keeping the posterior probability that the wrong hypothesis was accepted low, and may very well seek a sample size that would assure such a goal. This however is rather subtle; sample size determination

is a preexperimental exercise. What data will come our way is anybody's guess; posterior accuracies are by nature functions of the obtained data. There is a problem! One can and must therefore either seek an assurance of posterior accuracy for all possible data or at least all probable data: data that are likely to be seen once the prescribed sample size is used and data are obtained. This leads to very interesting mathematical problems. A moment's thinking will no doubt make us realize that the ability to simultaneously protect against all possible data by choosing a large sample corresponds mathematically to the uniform convergence to zero (uniform over the sample space, as the sample size goes to infinity) of an appropriate function of the observations $x_1, x_2, \ldots, x_n$. In DasGupta and Mukhopadhyay(1992), this was called uniform robustness. It was pointed out in that article that in many common problems of inference, uniform robustness is an unattainable goal even for simple priors. On hindsight, this is not so surprising. However, one can (almost) always assure posterior accuracy with respect to all samples outside of a small set with an arbitrarily low (predictive) probability. If by chance the obtained data happen to be one of the samples we were not preprotected against, the automatic accuracy is invalidated. Of course, this may also mean the modeling was wrong: why did such unanticipated data even arise? The exact sample size usually is found by solving a moderately messy equation treating the sample size n as a variable. For instance, in DasGupta and Mukhopadhyay, such explicit Bayesian sample sizes were provided for testing for a normal mean assuming a normal prior: the point is the user uses his own parameters for the prior, and a computer code built for the purpose solves the relevant equation. Prior flexibility with respect to the parameters of the prior is a must and is easily achieved.

Several critical questions deserve thinking. Is it important to have flexibility in the form of the functional form of the prior as well? In principle, the answer has to be 'yes'. Who knows what prior is deemed appropriate in a given problem? But it is quite plainly a fact that the associated mathematics changes completely with a change in the functional form of the prior. How many different problems are we going to solve with different priors? How many different codes are we going to write? And then there is the issue that there are dozens of standard statistical inference problems; are we going to write different codes with different functional forms for all of these problems? And finally, is it self defeating? A user in the real world most probably already has his/her classical sample size prescription.

Confronted with prescriptions of different sample sizes for different kinds of priors, it is very likely that the reaction will be one of resignation. It therefore appears to be natural that in standard inference problems (which usually happen to be normal theory problems), one writes a code for Bayesian sample sizes only for conjugate priors: the code will allow flexibility in the hyperparameters of the prior to suit the need of the individual user.

After the initial article, DasGupta and Mukhopadhyay (1992), in which one sample testing, multivariate testing (Hotelling's T^2 problem) and confidence set construction and some other minor problems were considered, in a subsequent article, DasGupta and Vidakovic(1994) give codes for Bayesian sample sizes for the problem of one way ANOVA. The model is thus the following:

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \varepsilon_{ij} \stackrel{iid}{\sim} N(0, \sigma^2),\ 1 \le i \le k,\ 1 \le j \le n.$$

One wants to test H_0: $\tau_1 = \ldots = \tau_k = 0$.

The user's input are the following: parameters μ_0 and σ_1^2 for a normal prior on μ; parameter σ_2^2 for a common prior for $\tau_1, \ldots, \tau_k$; a number $\varepsilon > 0$ which gives the desired posterior accuracy (actually the accuracy is $1 - \varepsilon$;); a number $\delta > 0$ which is used to determine preexperimentally a small set of samples of probability δ outside of which the given accuracy is preguaranteed; and finally a prior probability π for the null hypothesis. The mathematics used to arrive at the equation that one needs to solve is distributional theory of complicated quadratic forms. There are so many freely floating parameters as was just described, that a comprehensive table is beyond anybody's imagination. A code was provided. The following is a typical table of sample sizes obtained from the code.

Example 17. For the comparison of $k = 3$ treatments, with $\frac{\sigma_i^2}{\sigma^2} = 1$, the following sample sizes are needed to ensure a posterior accuracy of $1 - \varepsilon = .9$ outside of a set of predictive probability of .1, if an a priori probability of .5 is given to the null hypothesis of no effect.

σ^2	.5	.7	1	1.5	2
n	32	45	64	95	127

It is noticeable that they seem to be on the high side in comparison to standard sample sizes available in texts and books. Is this surprising? On hindsight again, it is not. The classical accuracy is an average accuracy. In the posterior Bayesian formulation of the

sample size problem, one is asking for guarantee of accuracy simultaneously against all but a minor set of samples. Furthermore, in the Bayesian formulation, one has uncertainties at two levels: the data, AND the parameter. A combination of all of these result in larger sample size prescriptions than the classical ones.

However, classical sample sizes have been around for a long time and will be around for a long time into the future. It is a natural curiosity and question of simple pragmatism to ask if adoption of the classical sample size will provide any kind of posterior accuracy one can at all live with. If in a series of problems, evidence accumulates that classical sample sizes are not really too bad as Bayesian sample sizes, it would be, realistically and pragmatically, time to stop solving complicated mathematics problems and writing involved mathematical codes for Bayesian sample sizes. The following is an example.

Example 18. We borrow an example from Montgomery (1984), in which it is wanted to test if the percentage of cotton in a fiber has an effect on its tensile strength. 5 different levels of cotton percentages are used and Montgomery evaluates the power of the standard F test at a 5% level when the difference between some pair of means is one standard deviation or more.

The following was found :

n	Power
40	.8
50	.86
60	.93

The goal is to find the implication to a Bayesian of using these sample sizes. Assuming an a priori probability of .5 for no effect (although it sounds as if in this problem, it should NOT be .5), let us see the value of δ implied by a specified accuracy $1 - \varepsilon = .95$. We take σ_1^2 and σ_2^2 to be equal and to be 1.33 times the error variance. Then the value of δ is respectively equal to .131, .095 and .073 for $n = 40$, 50 and 60 respectively. The very nature of the question we are asking is such that no cut and dried formulation or a cut and dried answer is possible. It seems as though that a guarantee of very good posterior accuracy can be given for most data by adoption of classical sample sizes geared towards a reasonable classical goal.

The problem of determining Bayesian sample sizes has just begun to be looked at. Standard problems, for most of which classical solutions are already available, do need to be looked at, at least for a while. We need to understand if new research is generally necessary, or classical solutions are fine. One final comment: the Bayesian problem could also be formulated as an average accuracy problem. This will no doubt lead to smaller required sample sizes; but to the strict Bayesian, such a sample size is not the correct one to ask for. And, of course, there is the other side of the coin: if a classical practitioner used the Bayesian sample size, what classical accuracy (power, etc.) s/he would be assured of? This is transparent in the problems addressed so far: Bayesian sample sizes are typically larger. The classicist would get better accuracy than usual with Bayesian sample sizes.

6. CONCLUDING REMARKS.

For various reasons, a behavioral evaluation of subjective prior Bayes rules is important or interesting or both. Without a careful evaluation of the soundness of a method, there is the clear danger of falling into the very tempting trap of writing a convenient set of assumptions. The danger, ironically, is more now with the advancement of computing technology. It is very easy to start thinking as if computers can replace thinking,introspection,and the human mind. In addition to these, questions like "what is the distribution of the Bayes estimate" are questions of basic and fundamental scientific curiosity. The difficulty of an answer only enhances the interest. There is a sea full of models and problems where such investigations ought to be done for a variety of reasons: synthesis of the two major paradigms is a good enough reason. Synthesis of Bayes and nonBayes statistics need not be looked at with cynicism, suspicion or ridicule. Whenever different schools of thought find, accidentally or through long and careful work, that significant common grounds exist, science progresses. It is a positive step, a step forward. We hope this article does a little to make a positive step.

Acknowledgements: J.K. Ghosh graciously read part of the manuscript and made some valuable comments. Saurabh Mukhopadhyay and Brani Vidakovic did the numerical calculations. I am delighted to thank both of them.

BIBLIOGRAPHY

Berger, J. (1985). Statistical decision theory and Bayesian analysis.

Bickel, P.J. and Blackwell, D. (1967). A note on unbiased estimates.

Bickel, P.J. and Lehmann, E.L. (1969). Unbiased estimation in convex families.

Bickel, P.J. (1984). Parametric robustness: small biases can be worthwhile.

Brockwell, P.J. and Davis, R.A. (1991). Time series: theory and methods.

Brown, L.D. (1971). Admissible estimators, recurrent diffusions and insoluble boundary value problems.

Brown, P.J. and Zidek, J.V. (1980). Multivariate adaptive ridge regression.

Casella, G. and Berger, R. (1987). Reconciling Bayesian and frequentist evidence in one sided testing problems.

DasGupta, A. (1994). Exact and approximate computation of expectations: with applications.

and Mukhopadhyay (1992). Uniform and subuniform posterior robustness and the sample size problem.

and Mukhopadhyay (1993). Uniform approximation of Bayes solutions and posteriors: frequentistly valid Bayes inference.

, McCabe, G. and Mukhopadhyay, S. (1994). Learning about an unknown proportion in mixed populations.

and Shyamalkumar, N.D. (1994). Bias and bias correction of Bayes estimates.

and Vidakovic, B. (1994). Sample sizes in one way ANOVA: the Bayesian viewpoint.

Diaconis, P. and Freedman, D. (1983). Frequency properties of Bayes rules.

Diaconis, P. (1985). Bayesian statistics as honest work.

Diaconis, P. (1986). Bayesian numerical analysis.

Efron, B. (1982). Maximum likelihood and decision theory.

Efron, B. and Tibshirani, R. (1993). An introduction to the bootstrap.

Erdelyi, A. (1955). Asymptotic expansions.

Gelfand, A. and Smith A.F.M. (1990). Sampling based approaches to calculating marginal densities.

Ghosh, J.K. and Mukherjee, R. (1991). Characterization of priors under which Bayesian and frequentist Bartlett corrections are equivalent.

Hartigan, J. (1965). An asymptotically unbiased prior distribution.

Hayman, W.K., Kershaw, D. and Lyons, T.J. (1984). Best harmonic approximant to a continuous function.

Hinkley, D.V. (1983). Can frequentist inference be very wrong? A conditional 'yes'.

Huber, P.J. (1981). Robust statistics.

Korner, T. (1989). Fourier analysis.

Lehmann, E.L. (1983). Theory of point estimation.

Lindley, D.V. (1988). The present position in Bayesian statistics: Wald memorial lecture.

Montgomery, D. (1984). Design and analysis of experiments.

Odeh, R.E. and Fox, M. (1975). Sample size choice.

Powell, M.J.D. (1981). Approximation theory and methods.

Reid, N. (1988). Saddlepoint methods.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations.

Sen, P.K. and Singer, J.M. (1993). Large sample methods in statistics: an introduction with applications.

Stein, C. (1985). On the coverage probability of confidence sets based on a prior distribution.

Tierney L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and

marginal densities.

Tukey, J.W. (1958). Bias and confidence in not-quite large samples.

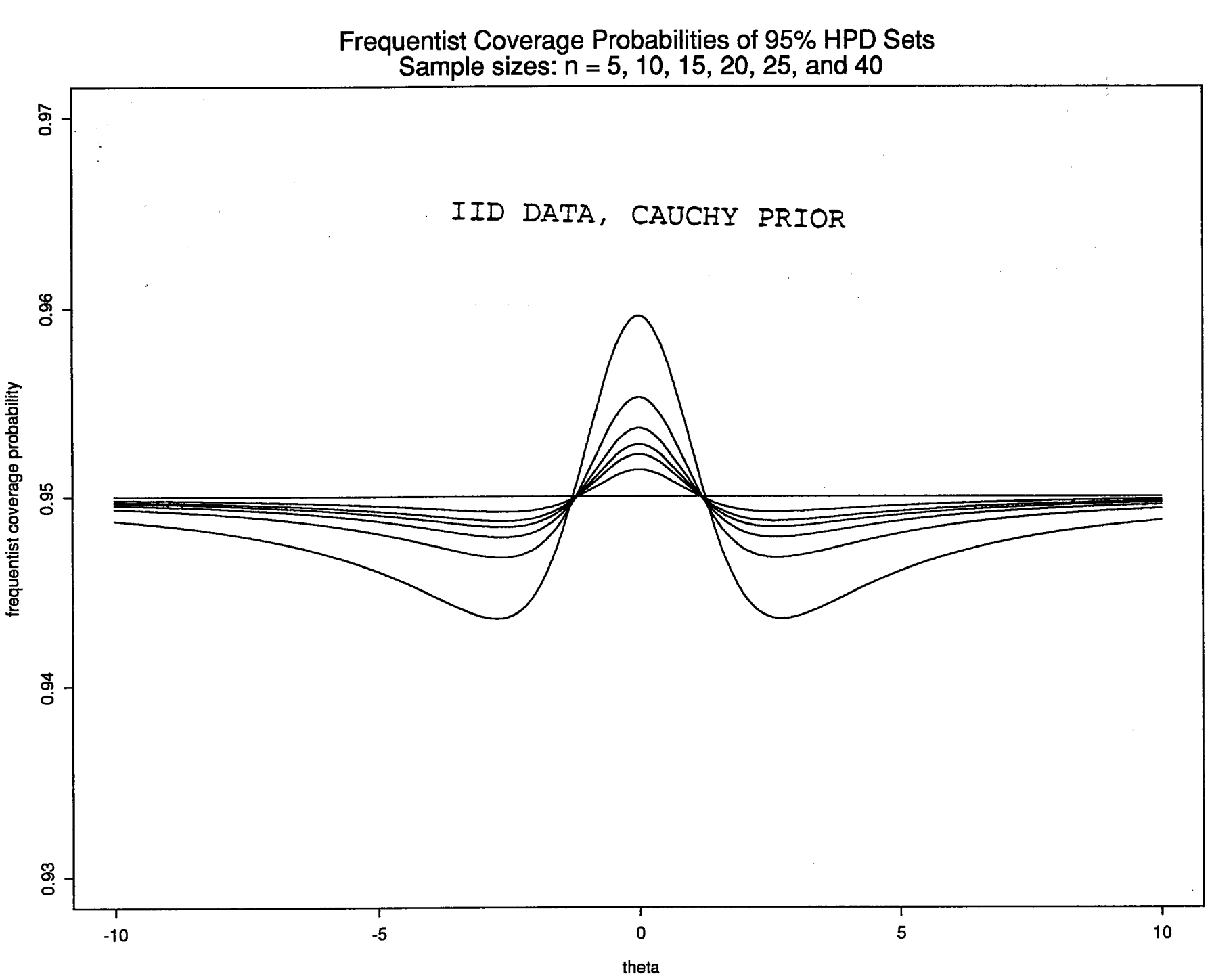
Unni, K. (1978). The theory of estimation in algebraic and analytic exponential families with applications to variance components models.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression.

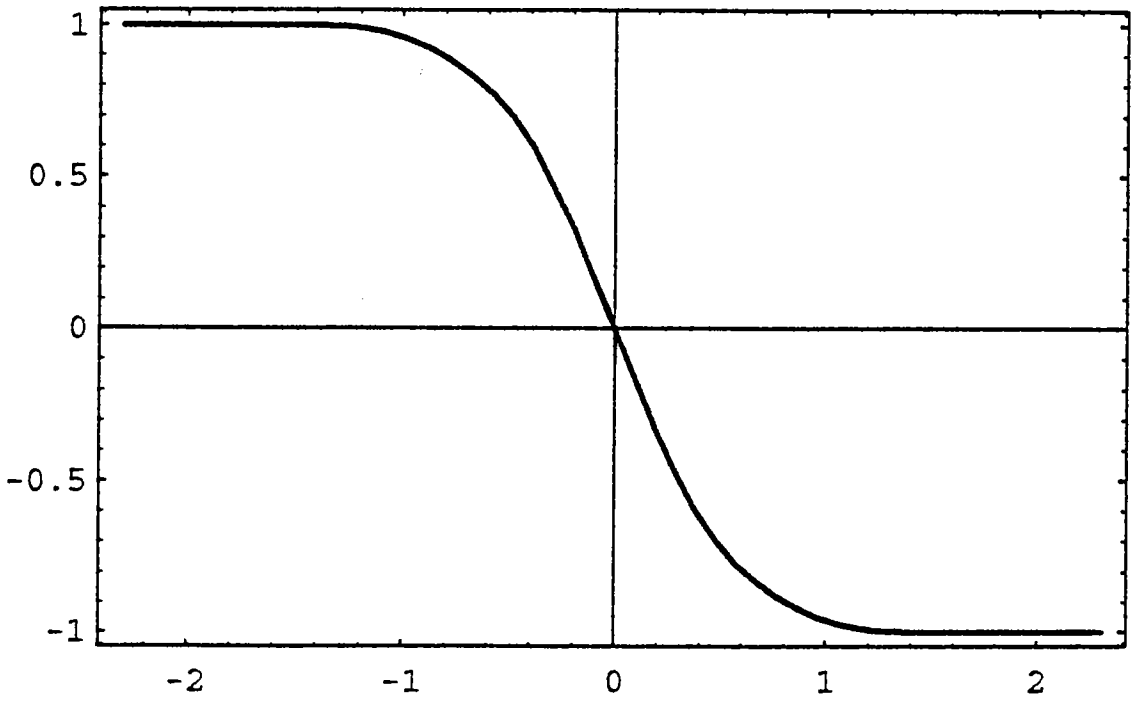
Wald, A. (1947). Sequential analysis.

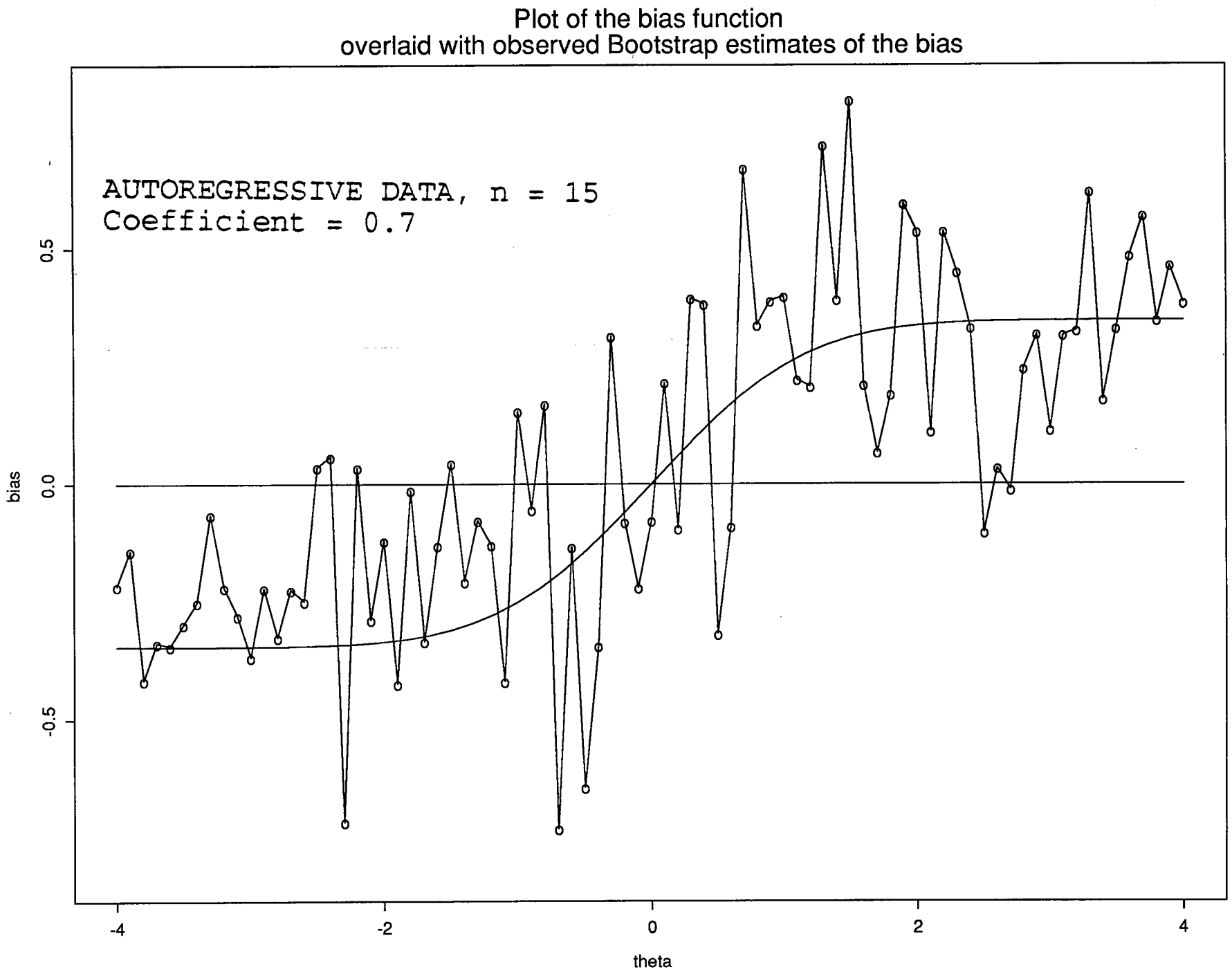
Wolfram, S. (1992). Mathematica - a system for doing mathematics by computer.

Woodroofe, M. (1976). Frequentist properties of Bayes sequential tests.


Frequentist Coverage Probabilities of 95% HPD Sets
Sample sizes: n = 5, 10, 15, 20, 25, and 40
IID DATA, CAUCHY PRIOR
frequentist coverage probability
0.97
0.96
0.95
0.94
0.93
-10
-5
0
5
10
theta

EXACT BIAS OF THE BAYES ESTIMATE :
IID DATA, n = 10, DOUBLE EXPONENTIAL
PRIOR




Plot of the bias function
overlaid with observed Bootstrap estimates of the bias
AUTOREGRESSIVE DATA, n = 15
Coefficient = 0.7
bias
0.5
0.0
-0.5
-4
-2
0
2
4
theta