Measuring and Controlling Bias for Some Bayesian Inferences and the Relation to Frequentist Criteria

Michael Evans and Yang Guo
Department of Statistical Sciences
University of Toronto

Abstract

A common concern with Bayesian methodology in scientific contexts is that inferences can be heavily influenced by subjective biases. As presented here, there are two types of bias for some quantity of interest: bias against and bias in favor. Based upon the principle of evidence, it is shown how to measure and control these biases for both hypothesis assessment and estimation problems. Optimality results are established for the principle of evidence as the basis of the approach to these problems. A close relationship is established between measuring bias in Bayesian inferences and frequentist properties that hold for any proper prior. This leads to a possible resolution to an apparent conflict between these approaches to statistical reasoning. Frequentism is seen as establishing a figure of merit for a statistical study, while Bayesianism plays the key role in determining inferences based upon statistical evidence.

Keywords and phrases: principle of evidence, bias against, bias in favor, plausible region, frequentism, confidence.

1 Introduction

A serious concern with Bayesian methodology is that the choice of the prior could result in conclusions that to some degree are predetermined before seeing the data. In certain circumstances this is correct. This can be seen by considering the problem associated with what is known as the Jeffreys-Lindley paradox where posterior probabilities of hypotheses, as well as associated Bayes factors, will produce increasing support for the hypothesis as the prior becomes more diffuse. So, while one may feel that a very diffuse prior is putting in very little information, it is in fact biasing the results in favor of the hypothesis. It has been argued, see Baskurt and Evans (2013) and Evans (2015), that the measurement and control of bias is a key element of a Bayesian analysis as without it, and the assurance that bias is minimal, the validity of any inference is suspect.

While attempts have been made to avoid the Jeffreys-Lindley paradox through the choice of the prior, modifying the prior to avoid bias is contrary to the ideals of a Bayesian analysis which requires the elicitation of a prior based upon knowledge of the phenomenon under study. Why should one change such a prior because of bias? Indeed, as will be discussed, there is bias in favor and bias against and typically choosing a prior to minimize one type of bias simply increases the other. The real method for controlling bias of both types is through the amount of data collected. So controlling bias is an aspect of design. Bias can be measured post-hoc and it then provides a way to assess the weight that should be given the results of an analysis. For example, if a study concludes that there is evidence in favor of a hypothesis, but it can be shown that there was a high prior probability that such evidence would be obtained, then the results of such an analysis can't be considered to be reliable.

Previous discussion concerning bias was focused on hypothesis assessment and in many ways this is a natural starting point. This paper is concerned with adding some aspects to those developments and to extending the approach to estimation and prediction problems. Furthermore, it is shown here that measuring and controlling bias establishes close links between a frequentist approach to statistics and Bayesian inference. In essence frequentism is concerned with design while inferences are Bayesian. Bayesian inference is based upon the evidence in the observed data and is unconcerned, at least for inference, about data sets that could have been obtained. Frequentism is concerned with the behavior of inferences as applied to unobserved data sets and this is entirely appropriate before the data is observed. So consideration of bias leads to a degree of unification between different ways of thinking about statistical reasoning.

The measurement of bias, and thus its control, is dependent upon measuring evidence. The *principle of evidence* is adopted here: evidence in favor of a specific value of an unknown occurs when the posterior probability of the value is greater than its prior probability, evidence against occurs when the posterior probability of the value is less than its prior probability and there is no evidence either way when these are equal. The major part of what is discussed here depends only on this simple principle but sometimes a numerical measure of evidence is needed and for this we use the *relative belief ratio* defined as the ratio of the posterior to prior probability. The relative belief ratio is related to the Bayes factor but has some nicer properties such as providing a measure of the evidence for each value of a parameter without the need to modify the prior.

There is not much discussion in the Bayesian literature of the notion of bias in the sense that is meant here. There is considerable discussion, however, concerning the Jeffreys-Lindley paradox and our position is that bias plays a key role in the issues that arise. Relevant recent papers on this include Shafer (1982), Spanos (2013), Sprenger (2013), Robert (2014), Cousins (2017) and Villa and Walker (2017) and these contain extensive background references. Gu et al. (2019) is concerned with the validation of quantum theory using Bayesian methodology applied to well-known data sets and the principle of evidence and an assessment of the bias in the prior plays a key role in the argument.

In Section 2 the concepts are defined, their properties are considered and

illustrated via a simple example where the Jeffreys-Lindley paradox is relevant. Also, it is seen that a well-known p-value does not satisfy the principle of evidence but can still be used to characterize evidence for or against but requires significance levels that go to 0 with increasing sample size or increasing diffuseness of the prior. In Section 3 the relationship with frequentism is discussed and a number of optimality results are established for the approach taken here to measuring and controlling bias, namely, via the principle of evidence. In Section 4, a variety of examples are considered and analyzed from the point-of-view of bias. All proofs of theorems are in the Appendix.

2   Evidence and Bias

For the discussion here there is a model $\{f_\theta : \theta \in \Theta\}$, given by densities f_θ, for data x and a proper prior probability distribution given by density π. It is supposed that interest is in inferences about $\psi = \Psi(\theta)$ where $\Psi : \Theta \to \Psi$ is onto and for economy the same notation is used for the function and its range. For the most part it is safe to assume all the probability distributions are discrete with results for the continuous case obtained by taking limits.

A measure of the evidence that $\psi \in \Psi$ is the true value is given by the relative belief ratio

$$RB_\Psi(\psi \,|\, x) = \lim_{\delta \to 0} \frac{\Pi_\Psi(N_\delta(\psi) \,|\, x)}{\Pi_\Psi(N_\delta(\psi))} = \frac{\pi_\Psi(\psi \,|\, x)}{\pi_\Psi(\psi)} \tag{1}$$

where $\Pi_\Psi, \Pi_\Psi(\cdot \,|\, x)$ are the prior and posterior probability measures of Ψ with densities π_Ψ and $\pi_\Psi(\cdot \,|\, x)$, respectively, and $N_\delta(\psi)$ is a sequence of sets converging nicely to $\{\psi\}$. The last equality in (1) requires some conditions but the prior density positive and continuous at ψ is enough. So $RB_\Psi(\psi \,|\, x) > 1$ implies evidence for the true value being ψ, etc. Any *valid* measure of evidence should satisfy the principle of evidence, namely, the existence of a cut-off value that determines evidence for and against as prescribed by the principle. Naturally, this cut-off is 1 for the relative belief ratio. The Bayes factor is also a valid measure of evidence and with the same cut-off. When $\Pi_\Psi(A) > 0$ then the Bayes factor of A equals $RB(A \,|\, x)/RB(A^c \,|\, x)$ and so can be defined in terms of the relative belief ratio, but not conversely. Also, $RB(A \,|\, x) > 1$ iff $RB(A^c \,|\, x) < 1$ and so the Bayes factor is not really a comparison of the evidence for A being true with the evidence for its negation. In the continuous case, if we define the Bayes factor for ψ as a limit as in (1), then this limit equals $RB_\Psi(\psi \,|\, x)$. Further discussion on the choice of a measure of evidence can be found in Evans (2015) as there are other candidates beyond these two. It is important to note, however, that the discussion of bias depends only on the principle of evidence and is the same no matter what valid measure of evidence is used.

The following example is carried along as it illustrates a number of things.

Example 1. *Location normal.*

Suppose $x = (x_1, \ldots, x_n)$ is i.i.d. $N(\mu, \sigma_0^2)$ with π a $N(\mu_0, \tau_0^2)$ prior. Then

$$\mu \,|\, x \sim N\big((n/\sigma_0^2 + 1/\tau_0^2)^{-1}\,(n\bar{x}/\sigma_0^2 + \mu_0/\tau_0^2),\, (n/\sigma_0^2 + 1/\tau_0^2)^{-1}\big) \text{ so}$$

$$RB(\mu \,|\, x) = \left(1 + \frac{n\tau_0^2}{\sigma_0^2}\right)^{1/2} \exp\left\{ \begin{array}{c} -\frac{1}{2}\left(1 + \frac{\sigma_0^2}{n\tau_0^2}\right)\left(\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma_0} + \frac{\sigma_0(\mu_0-\mu)}{\sqrt{n}\tau_0^2}\right)^2 \\ + \frac{(\mu-\mu_0)^2}{2\tau_0^2} \end{array} \right\}.$$

2.1 Bias in Hypothesis Assessment Problems

The value $RB_\Psi(\psi_* \,|\, x)$ tells us if we have evidence for or against $H_0 : \Psi(\theta) = \psi_*$.

Example 1. *Location normal (continued).*

Observe that as $\tau_0^2 \to \infty$, then $RB(\mu \,|\, x) \to \infty$ for every μ and in particular for a hypothesized value $H_0 = \{\mu_*\}$. So it would appear that overwhelming evidence is obtained for the hypothesis when the prior is very diffuse and this holds irrespective of what the data says. Also, when the standardized value $\sqrt{n}|\bar{x} - \mu_*|$ is fixed, then $RB(\mu_* \,|\, x) \to \infty$ as $n \to \infty$. This phenomenon also occurs if a Bayes factor (which equals $RB(\mu_* \,|\, x)$ in this case) or a posterior probability based upon a discrete prior mass at μ_* is used to assess H_0. Accordingly all these measures lead to a sharp disagreement with the frequentist p-value $2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_*|/\sigma_0))$ when it is small. This is the Jeffreys-Lindley paradox and it arises quite generally.

The Jeffreys-Lindley paradox shows that the strength of evidence cannot be measured strictly by the size of the measure of evidence. A logical way to assess this is to compare the evidence for ψ_* with the evidence for the other possible values for ψ. The *strength* of the evidence can then be measured by

$$\Pi_\Psi(RB_\Psi(\psi \,|\, x) \leq RB_\Psi(\psi_* \,|\, x) \,|\, x), \tag{2}$$

the posterior probability that the true value has evidence no greater than the evidence for ψ_*. So if $RB_\Psi(\psi_* \,|\, x) < 1$ and (2) is small, then there is strong evidence against ψ_* while, if $RB_\Psi(\psi_* \,|\, x) > 1$ and (2) is large, then there is strong evidence in favor of ψ_*. The inequalities $\Pi_\Psi(\{\psi_*\} \,|\, x) \leq \Pi_\Psi(RB_\Psi(\psi \,|\, x) \leq RB_\Psi(\psi_* \,|\, x) \,|\, x) \leq RB_\Psi(\psi_* \,|\, x)$ hold and so when $RB_\Psi(\psi_* \,|\, x)$ is small there is strong evidence against ψ_* and when $RB_\Psi(\psi_* \,|\, x) > 1$ and $\Pi_\Psi(\{\psi_*\} \,|\, x)$ is big, then there is strong evidence in favor of ψ_*. Note, however, that $\Pi_\Psi(\{\psi_*\} \,|\, x) \approx 1$ does not guarantee $RB_\Psi(\psi_* \,|\, x) > 1$ and if $RB_\Psi(\psi_* \,|\, x) < 1$ this means that there is weak evidence against ψ_*. There is no reason why multiple measures of the strength of the evidence can't be used (see the discussion in Section 2.2). There are some issues with (2) in the continuous case that require a modification and we refer to Evans (2015) for this as the strength does not play a key role in the discussion here. The important point is to somehow calibrate the measure of evidence using probability to measure how strong belief in the evidence is.

Example 1. *Location normal (continued).*

A simple calculation shows that, with $\sqrt{n}|\bar{x}-\mu_*|$ fixed, then (2) converges to $2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_*|/\sigma_0))$ as $n\tau_0^2 \to \infty$. So, if the p-value is small, this indicates that a large value of $RB_\Psi(\mu_* \,|\, x)$ is only weak evidence in favor of μ_*. It is to be noted that the p-value $2(1-\Phi(\sqrt{n}|\bar{x}-\mu_*|/\sigma_0))$ is not a valid measure of evidence

as described here because there is no cut-off that corresponds to evidence for and evidence against. So its appearance as a measure of the strength of the evidence is not in any sense circular.

Simple algebra shows, however, that $2(1-\Phi(\sqrt{n}|\bar{x}-\mu_*|/\sigma_0))-2(1-\Phi([\log(1+n\tau_0^2/\sigma_0^2) + (1 + n\tau_0^2/\sigma_0^2)^{-1}(\bar{x}-\mu_0)^2/\tau_0^2]^{1/2})$, a difference of two p-values, is a valid measure of evidence via the cut-off 0. From this it is seen that the values of the first p-value $2(1 - \Phi(\sqrt{n}|\bar{x} - \mu_*|/\sigma_0)$ that lead to evidence against generally become smaller as $n\tau_0^2 \to \infty$. For example, with $n = 10, \sigma_0^2 = 1, \mu_* = 0$ and $\sqrt{n}|\bar{x} - \mu_*| = 1.96$, then the p-value equals 0.05. Setting $\mu_0 = 0$ and $\tau_0^2 = 1$ the second p-value equals 0.119 and so there is evidence against, with $\tau_0^2 = 10$ the the second term equals 0.032 and with $\tau_0^2 = 100$ it equals 0.009, so there is evidence in favor in both cases. When n increases these values become smaller as with $n = 50$ the first p-value equal to 0.05 is always evidence in favor. Similar results are obtained with a uniform prior on $(-m, m)$, reflecting perhaps a desire to treat many values equivalently, as $m \to \infty$ or $n \to \infty$. For example, with $m = 10$ and $n = 10, \sigma_0^2 = 1, \mu_* = 0, \sqrt{n}|\bar{x} - \mu_*| = 1.96$, then the second p-value equals 0.002 and there is evidence in favor. These conclusions are similar to those found in Berger and Selke (1987) and Berger and Delampady (1987).

It is very simple to elicit (μ_0, τ_0^2) based on prescribing an interval that contains the true μ with some high probability such as 99.9%, taking μ_0 to be the mid-point and so τ_0^2 is determined. There is no reason to take τ_0^2 to be arbitrarily large. But still one wonders if the choice made is inducing some kind of bias into the problem as taking τ_0^2 too large clearly does.

Certainly default choices of priors should be avoided when possible, but even when eliciting, how can we know if the chosen prior is inducing bias? To assess this a numerical measure is required. The principle of evidence suggests that *bias against* H_0 is measured by

$$M(RB_\Psi(\psi_* \,|\, X) \le 1 \,|\, \psi_*) \tag{3}$$

where $M(\cdot \,|\, \psi_*)$ is the prior predictive distribution of the data given that the hypothesis is true. So (3) is the prior probability that evidence in favor of ψ_* will not be obtained when ψ_* is the true value. If (3) is large, then there is an *a priori* bias against H_0.

For the bias in favor of H_0 it is necessary to assess if evidence against H_0 will not be obtained with high prior probability even when H_0 is false. One possibility is to measure *bias in favor* by

$$\int_{\Psi\backslash\{\psi_*\}} M(RB_\Psi(\psi_* \,|\, X) \ge 1 \,|\, \psi) \, \Pi_\Psi(d\psi)$$
$$= M(RB_\Psi(\psi_* \,|\, X) \ge 1) - M(RB_\Psi(\psi_* \,|\, X) \ge 1 \,|\, \psi_*)\Pi_\Psi(\{\psi_*\}) \tag{4}$$

which is the prior probability of not obtaining evidence against ψ_* when it is false. When $\Pi_\Psi(\{\psi_*\}) = 0$, then (4) equals $M(RB_\Psi(\psi_* \,|\, X) \ge 1)$ where M is the prior predictive for the data. For continuous parameters it can be argued that it does not make sense to consider values of ψ so close to ψ_* that they

are practically speaking indistinguishable. Suppose then there is a measure of distance d_Ψ on Ψ and a value $\delta > 0$ such that, if $d_\Psi(\psi_*, \psi) < \delta$, then ψ_* and ψ are indistinguishable in the application. The *bias in favor* of H_0 can then be measured by replacing $\Psi \backslash \{\psi_*\}$ in (4) by $\{\psi : d_\Psi(\psi_*, \psi) \geq \delta\}$ which has upper bound bound

$$\sup_{\psi : d_\Psi(\psi_*, \psi) \geq \delta} M(RB_\Psi(\psi_* \,|\, X) \geq 1 \,|\, \psi). \qquad (5)$$

Typically $M(RB_\Psi(\psi_* \,|\, X) \geq 1 \,|\, \psi)$ decreases as ψ moves away from ψ_* so (5) can be computed by finding the supremum over the set $\{\psi : d_\Psi(\psi_*, \psi) = \delta\}$ and, when ψ is real-valued and d_Ψ is Euclidian distance, this equals $\{\psi_* - \delta, \psi_* + \delta\}$.

It is to be noted that the measures of bias given by (3), (4) and (5) do not depend on using the relative belief ratio to measure evidence. Any valid measure of evidence will determine the same values when the relevant cut-off is substituted for 1. It is only (2) that depends on the specific choice of the relative belief ratio as the measure of evidence.

Under general circumstances, see Evans (2015), both biases will converge to 0 as the amount of data increases and so both biases can be controlled by design. Clearly there is no point in reporting the results of an analysis when there is a lot of bias unless the evidence actually contradicts the bias.

Example 1. *Location normal (continued).*

Under $M(\cdot \,|\, \mu)$, then $\bar{x} \sim N(\mu, \tau_0^2 + \sigma_0^2/n)$. So, putting

$$a(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n) = \sigma_0(\mu_* - \mu_0)/\sqrt{n}\tau_0^2,$$

$$b(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n) = \left\{ \left(1 + \frac{\sigma_0^2}{n\tau_0^2}\right) \left[\log\left(1 + \frac{n\tau_0^2}{\sigma_0^2}\right) + \frac{(\mu_* - \mu_0)^2}{\tau_0^2}\right] \right\}^{1/2},$$

then (3) is given by

$$M(RB(\mu_* \,|\, X) \leq 1 \,|\, \mu_*) = 1 - \Phi\left(a(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n) + b(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n)\right) +$$
$$\Phi\left(a(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n) - b(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n)\right). \quad (6)$$

This goes to 0 as $n \to \infty$ or as $\tau_0^2 \to \infty$. So bias against can be controlled by sample size n or by the diffuseness of the prior although, as subsequently shown, a diffuse prior induces bias in favor. It is also the case that (6) converges to 0 when $\mu_0 \to \pm\infty$ or when $\sigma_0/\sqrt{n}\tau_0$ is fixed and $\tau_0 \to 0$. So it would appear that using a prior with location quite different than the hypothesized value or a prior that was much more concentrated than the sampling distribution, can be used to lower bias against. These are situations, however, where one can expect to have prior-data conflict after observing the data.

The entries in Table 1 record the bias against for a specific case and illustrate that increasing n does indeed reduce bias. The entries also show that bias against can be greater when the prior is centered on the hypothesis. Figure 1 contains a plot of the bias against $H_0 = \{\mu_*\}$, as a function of μ_*, when using a $N(0,1)$ prior. Note that the maximum bias against occurs at the mean of the prior (and equals 0.143) and this typically occurs when $\sigma_0^2/n\tau_0^2 < 1$,

n	$\mu_0 = 1, \tau_0 = 1$	$\mu_0 = 0, \tau_0 = 1$
5	0.095	0.143
10	0.065	0.104
20	0.044	0.074
50	0.026	0.045
100	0.018	0.031

Table 1: Bias against for the hypothesis $H_0 = \{0\}$ with a $N(\mu_0, \tau_0^2)$ prior for different sample sizes n with $\sigma_0 = 1$.
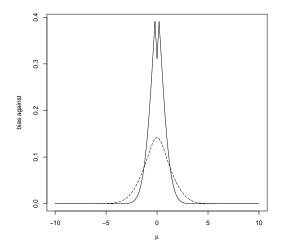

Figure 1: Plot of bias against $H_0 = \{\mu\}$ with a $N(0,1)$ prior (- - -) and a $N(0, 0.01)$ prior (——) with $n = 5, \sigma_0 = 1$.

namely, when the data is more concentrated than the prior. Figure 1 also contains a plot of the bias against when using a prior more concentrated than the data distribution. That the bias against is maximized, as a function of the hypothesized mean μ_*, when μ_* equals the value associated with the strongest belief under the prior seems odd. This phenomenon arises quite often, and the mathematical explanation for this is that the greater the amount of prior probability assigned to a value, the harder it is for the posterior probability to increase and so it is quite logical when considering evidence. It will be seen that this phenomenon is very convenient for the control of bias in estimation problems and could be used as an argument for using a prior centered on the hypothesis, although this is not necessary as beliefs may be different.

Now consider (5), namely, bias in favor of $H_0 = \{\mu_*\}$. Putting

$$c(\mu_*, \mu, \mu_0, \tau_0^2, \sigma_0^2, n) = \sqrt{n}(\mu_* - \mu)/\sigma_0 + a(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n),$$
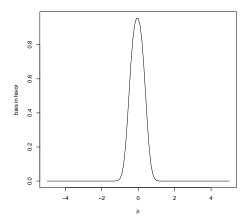
7

Figure 2: Plot of $M(RB(0\,|\,X) \geq 1\,|\,\mu)$ when $n = 20, \mu_0 = 1, \tau_0 = 1, \sigma_0 = 1$.

then (5) equals $\max M(RB(\mu_* \,|\, X) \geq 1 \,|\, \mu_* \pm \delta)$ where

$$M(RB(\mu_* \,|\, X) \geq 1 \,|\, \mu) = \Phi\left(c(\mu_*, \mu, \mu_0, \tau_0^2, \sigma_0^2, n) + b(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n)\right) -$$
$$\Phi\left(c(\mu_*, \mu, \mu_0, \tau_0^2, \sigma_0^2, n) - b(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n)\right) \quad (7)$$

which converges to 0 as $n \to \infty$ and also as $\mu \to \pm\infty$. But (7) converges to 1 as $\tau_0^2 \to \infty$, so if the prior is too diffuse there will be bias in favor of μ_*. So resolving the Jeffreys-Lindley paradox requires choosing the sample size n, after choosing the prior, so that (7) is suitably small. Note that choosing τ_0^2 larger reduces bias against but increases bias in favor and so generally bias cannot be avoided by choice of prior. Figure 2 is a plot of $M(RB(\mu_* \,|\, X) \geq 1 \,|\, \mu)$ for a particular case and this strictly decreases as μ moves away from μ_*.

In Table 2 we have recorded some specific values of the bias in favor using (4) and using (5) where d_Ψ is Euclidean distance. It is seen that bias in favor can be quite serious for small samples. When using (5) this can be mitigated by making δ larger. For example, with $(\mu_0, \tau_0) = (0, 1), \delta = 1.0, n = 20$ the bias in favor equals 0.004. Note, however, that δ is not chosen to make the bias in favor small, rather it is determined in an application as the difference from the null that is just practically important. The virtues of determining a suitable value of δ are also readily apparent as (5) is much smaller than (4) for larger n.

A comparison of Tables 1 and 2 shows that a study whose purpose is to demonstrate evidence in favor of H_0 is much more demanding than one whose purpose is to determine whether or not there is evidence against H_0.

n	$(\mu_0, \tau_0) = (1, 1)$	$(\mu_0, \tau_0) = (0, 1)$
5	0.323 (0.871)	0.451 (0.631)
10	0.259 (0.747)	0.371 (0.516)
20	0.215 (0.519)	0.299 (0.327)
50	0.153 (0.125)	0.219 (0.062)
100	0.116 (0.006)	0.168 (0.002)

Table 2: Bias in favor of the hypothesis $H_0 = \{0\}$ with a $N(\mu_0, \tau_0^2)$ prior for different sample sizes n with $\sigma_0 = 1$ using (4) (and using (5) with $\delta = 0.5$).

2.2 Bias in Estimation Problems

The relative belief estimate of $\psi = \Psi(\theta)$ is the value that maximizes the measure of evidence, namely, $\psi(x) = \arg\sup RB_\Psi(\psi \,|\, x)$. It is easy to show that $RB_\Psi(\psi(x) \,|\, x) \geq 1$ with the inequality strict except in trivial contexts. The accuracy of this estimate can be measured by the "size" of the *plausible region* $Pl_\Psi(x) = \{\psi : RB_\Psi(\psi \,|\, x) > 1\}$, the set of values of ψ that have evidence in their favor and note $\psi(x) \in Pl_\Psi(x)$. To say that $\psi(x)$ is an accurate estimate, requires that $Pl_\Psi(x)$ be "small", perhaps as measured by $Vol(Pl_\Psi(x))$ where Vol is some measure of volume, and also have high posterior content $\Pi_\Psi(Pl_\Psi(x) \,|\, x)$ which measures the belief that the true value is in $Pl_\Psi(x)$. Note that $Pl_\Psi(x)$ does not depend on the specific measure of evidence chosen, in this case the relative belief ratio. Any valid estimator must satisfy the principle of evidence and so be in $Pl_\Psi(x)$. It is argued that in an estimation problem, bias is measured by various coverage probabilities for the plausible region.

Note too that if there is evidence in favor of $H_0 : \Psi(\theta) = \psi_*$, then $\psi_* \in Pl_\Psi(x)$ and so represents the natural estimate of ψ provided there was a clear reason for assessing the evidence for this value. The strength of the evidence in favor of ψ_* can then also be measured by the size of $Pl_\Psi(x)$. Similarly, if evidence against H_0 is obtained then $\psi_* \in Im_\Psi = \{\psi : RB_\Psi(\psi \,|\, x) < 1\}$, the *implausible region* and then there is strong evidence against H_0 provided $Pl_\Psi(x)$ has small volume and large posterior probability. A virtue of this approach to measuring the strength of the evidence is that it does not depend upon using the relative belief ratio to measure evidence.

2.2.1 Bias Against

The prior probability that the plausible region does not cover the true value measures bias against when estimating ψ. For if this probability is large, then the estimate and the plausible region are *a priori* likely to be misleading as to the true value. The prior probability that $Pl_\Psi(x)$ doesn't contain $\psi = \Psi(\theta)$ when $\theta \sim \Pi, X \sim P_\theta$ is

$$E_{\Pi_\Psi}\left(M(\psi \notin Pl_\Psi(X) \,|\, \psi)\right) = E_{\Pi_\Psi}(M(RB_\Psi(\psi \,|\, X) \leq 1 \,|\, \psi)) \qquad (8)$$

which is also the average bias against over all hypothesis testing problems $H_0 : \Psi(\theta) = \psi$. Note that $1 - E_{\Pi_\Psi}\left(M(\psi \notin Pl_\Psi(X) \,|\, \psi)\right) = E_{\Pi_\Psi}\left(M(\psi \in Pl_\Psi(X) \,|\, \psi)\right)$

n	$\tau_0 = 1$	$\tau_0 = 0.5$
5	0.107	0.193
10	0.075	0.146
20	0.051	0.107
50	0.031	0.067
100	0.021	0.046

Table 3: Average bias against $H_0 = 0$ when using a$N(0, \tau_0^2)$ prior for different sample sizes n.

$= E_M \left(\Pi_\Psi (Pl_\Psi(X) \,|\, X) \right)$ which is the prior coverage probability of Pl_Ψ. Also,

$$\sup_\psi M(\psi \notin Pl_\Psi(X) \,|\, \psi) = \sup_\psi M(RB_\Psi(\psi \,|\, X) \leq 1 \,|\, \psi), \qquad (9)$$

is an upper bound on (8). Therefore, controlling (9) controls the bias against in estimation and all hypothesis assessment problems involving ψ. Also $1 - \sup_\psi M(\psi \notin Pl_\Psi(X) \,|\, \psi) = \inf_\psi M(\psi \in Pl_\Psi(X) \,|\, \psi) \leq E_M \left(\Pi_\Psi(Pl_\Psi(X) \,|\, X) \right)$ so using (9) implies lower bounds for the coverage probability and for the expected posterior content of the plausible region. In general, both (8) and (9) converge to 0 with increasing amounts of data. So it is possible to control for bias against in estimation problems by design.

Example 1. *Location normal (continued).*

The value of $M(RB(\mu \,|\, X) \leq 1 \,|\, \mu)$ is given in (6) and examples are plotted in Figure 1. When $\mu \sim N(\mu_0, \tau_0^2)$ then $z = (\mu - \mu_0)/\tau_0 \sim N(0, 1)$ so

$E_\Pi \left(M(RB(\mu \,|\, X) \leq 1 \,|\, \mu) \right)$

$$= 1 - E \left[\begin{array}{c} \Phi \left(\frac{\sigma_0}{\sqrt{n}\tau_0} Z + \left\{ \left(1 + \frac{\sigma_0^2}{n\tau_0^2} \right) \left[\log \left(1 + \frac{n\tau_0^2}{\sigma_0^2} \right) + Z^2 \right] \right\}^{1/2} \right) + \\ \Phi \left(\frac{\sigma_0}{\sqrt{n}\tau_0} Z - \left\{ \left(1 + \frac{\sigma_0^2}{n\tau_0^2} \right) \left[\log \left(1 + \frac{n\tau_0^2}{\sigma_0^2} \right) + Z^2 \right] \right\}^{1/2} \right) \end{array} \right]$$

which is notably independent of the prior mean μ_0. The dominated convergence theorem implies $E_\Pi \left(M(RB(\mu \,|\, X) \leq 1 \,|\, \mu) \right) \to 0$ as $n \to \infty$ or as $\tau_0^2 \to \infty$. So provided $n\tau_0^2/\sigma_0^2$ is large enough, there is no estimation bias against. Table 3 illustrates some values of this bias measure. Subtracting the probabilities in Table 3 from 1 gives the prior probability that the plausible region covers the true value and the expected posterior content of the plausible region. So when $n = 20, \tau_0 = 1$, the prior probability of the plausible region containing the true value is $1 - 0.051 = 0.949$ so $Pl(x)$ is a 0.949 Bayesian confidence interval for μ.

To use (9) it is necessary to maximize $M(RB(\mu \,|\, X) \leq 1 \,|\, \mu)$ as a function of μ and it is seen that, at least when the prior is not overly concentrated, that this maximum occurs at μ_0. Figure 1 shows that when using the $N(0, 1)$ prior the maximum occurs at $\mu = 0$ when $n = 5$ and from the second column of Table 1, the maximum equals 0.143. The average bias against is given by 0.107, as recorded in Table 3. Note that the maximum also occurs at $\mu = 0$ for the other values of n recorded in Table 1.

2.2.2 Bias in Favor

Bias in favor occurs when the prior probability that Im_Ψ does not cover a false value is large, namely, when

$$\int_\Psi \int_{\Psi \setminus \{\psi_*\}} M(\psi_* \notin Im_\Psi(X) \,|\, \psi) \, \Pi_\Psi(d\psi) \, \Pi_\Psi(d\psi_*)$$

$$= \int_\Psi \int_{\Psi \setminus \{\psi_*\}} M(RB_\Psi(\psi_* \,|\, X) \geq 1 \,|\, \psi) \, \Pi_\Psi(d\psi) \, \Pi_\Psi(d\psi_*) \qquad (10)$$

is large as this would seem to imply that the plausible region will cover a randomly selected false value from the prior with high prior probability. Note that (10) is the prior mean of (4) and in the continuous case equals $\int_\Psi M(\psi_* \notin Im_\Psi(X)) \, \Pi_\Psi(d\psi_*)$. As previously discussed, however, it often doesn't make sense to distinguish values of ψ that are close to ψ_*. The bias in favor for estimation can then be measured by

$$E_{\Pi_\Psi} \left(\sup_{\psi : d_\Psi(\psi, \psi_*) \geq \delta} M(\psi_* \notin Im_\Psi(X) \,|\, \psi) \right)$$

$$= E_{\Pi_\Psi} \left(\sup_{\psi : d_\Psi(\psi, \psi_*) \geq \delta} M(RB_\Psi(\psi_* \,|\, X) \geq 1 \,|\, \psi) \right). \qquad (11)$$

An upper bound on (11) is commonly equal to 1 as illustrated in Figure 3 and so is not useful.

It is the size and posterior content of $Pl_\Psi(x)$ that provides a measure of the accuracy of the estimate $\psi(x)$. As discussed in Section 2.2.1 the *a priori* expected posterior content of $Pl_\Psi(x)$ can be controlled by bias against. The *a priori* expected volume of $Pl_\Psi(x)$ satisfies

$$E_M \left(Vol(Pl_\Psi(X)) \right) = \int_\Psi \int_\Psi M(\psi_* \in Pl_\Psi(X) \,|\, \psi) \, \Pi_\Psi(d\psi) \, Vol(d\psi_*). \qquad (12)$$

Notice that when $\Pi_\Psi(\{\psi\}) = 0$ for every ψ, this can be interpreted as a kind of average of the prior probabilities of the plausible region covering a false value.

Example 1. *Location normal (continued).*
It follows from (7) that

$$\sup M(RB(\mu_* \,|\, X) \geq 1 \,|\, \mu_* \pm \delta)$$

$$= \sup \left\{ \begin{array}{l} \Phi\left(c(\mu_*, \mu_* \pm \delta, \mu_0, \tau_0^2, \sigma_0^2, n) + b(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n)\right) - \\ \Phi\left(c(\mu_*, \mu_* \pm \delta, \mu_0, \tau_0^2, \sigma_0^2, n) - b(\mu_*, \mu_0, \tau_0^2, \sigma_0^2, n)\right) \end{array} \right\}$$

Note that as $\mu_* \to \pm\infty$, then $M(RB(\mu_* \,|\, X) \geq 1 \,|\, \mu_* \pm \delta) \to 1$ when $n\tau_0^2/\sigma_0^2 > 1$, see Figure 3, and converges to 0 if $n\tau_0^2/\sigma_0^2 < 1$, so it would appear that the better circumstance for guarding against bias in favor is when the prior is putting in more information than the data. As previously noted, however, this is a situation where we might expect prior data-conflict to arise and, except in exceptional
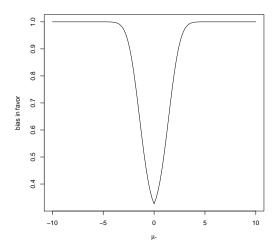
Figure 3: Bias in favor of μ maximized over $\mu \pm \delta$ based on a $N(0,1)$ prior and $\sigma_0 = 1, n = 20, \delta = 0.5$.

n	$(\mu_0, \tau_0) = (0,1), \delta = 1.0$	$(\mu_0, \tau_0) = (0,1), \delta = 0.5$
5	0.451	0.798
10	0.185	0.690
20	0.025	0.486
50	0.000	0.131
100	0.000	0.009

Table 4: Average bias in favor for estimation based on (11) when using a $N(0, \tau_0^2)$ prior for different sample sizes n and difference δ.

n	$\tau_0 = 1$	$\tau_0 = 0.5$
5	0.625 (0.893)	0.491 (0.807)
10	0.499 (0.925)	0.389 (0.854)
20	0.393 (0.949)	0.312 (0.893)
50	0.281 (0.969)	0.231 (0.933)
100	0.215 (0.979)	0.181 (0.954)

Table 5: Expected half-widths (coverages) of the plausible interval when using a $N(\mu_0, \tau_0^2)$ prior for different sample sizes n.

circumstances should be avoided. Table 4 contains values of (7) for this situation with different values of δ.

Some elementary calculations give $Pl(x) = \bar{x} \pm w(\bar{x}, n, \sigma_0^2, \mu_0, \tau_0^2)$ with

$$
w(\bar{x}, n, \sigma_0^2, \mu_0, \tau_0^2)
$$

$$
= \frac{\sigma_0}{\sqrt{n}} \left(1 + \frac{n\tau_0^2}{\sigma_0^2}\right)^{-1/2} \left\{ \left(1 + \frac{n\tau_0^2}{\sigma_0^2}\right) \log\left(1 + \frac{n\tau_0^2}{\sigma_0^2}\right) + \left(\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}}\right)^2 \right\}^{1/2}
$$

where $z = \sqrt{n}(\bar{x} - \mu_0)/\sigma_0 \sim N(0, 1)$ under M. It is notable that the prior distribution of the width is independent of the prior mean. Table 5 contains some expected half-widths together with the coverage probabilities of $Pl(x)$.

3 Frequentist and Optimal Properties

Consider now the bias against $H_0 = \{\psi_*\}$, namely, $M(RB_\Psi(\psi_* \mid X) \leq 1 \mid \psi_*)$. If we repeatedly generate $\theta \sim \pi(\cdot \mid \psi_*), X \sim f_\theta$, then this probability is the long-run proportion of times that $RB_\Psi(\psi_* \mid X) \leq 1$. This frequentist interpretation depends on the conditional prior $\pi(\cdot \mid \psi_*)$ and when $\Psi(\theta) = \theta$, so there are no nuisance parameters, this is a "pure" frequentist probability. Even in the latter case there is some dependence on the prior, however, as $RB(\theta_* \mid x) = f_{\theta_*}(x)/m(x)$ so x satisfies $RB_\Psi(\theta_* \mid x) \leq 1$ iff $f_{\theta_*}(x) \leq m(x)$ where $m(x) = \int_\Theta f_\theta(x) \Pi(d\theta)$. So in general the region $\{x : RB_\Psi(\psi_* \mid x) \leq 1\}$ depends on π but the probability $M(RB_\Psi(\psi_* \mid X) \leq 1 \mid \psi_*)$ depends only on the conditional prior predictive given $\Psi(\theta) = \psi_*$, namely, $m(x \mid \psi_*) = \int_\Theta f_\theta(x) \Pi(d\theta \mid \psi_*)$, and not on the marginal prior π_Ψ on ψ. We refer to probabilities that depend only on $M(\cdot \mid \psi_*)$ as frequentist, for example, coverage probabilities are called confidences, and those that depend on the full prior π as Bayesian confidences. The frequentist label is similar to use of the confidence terminology when dealing with random effects models as nuisance parameters have been integrated out.

Suppose now that some other general rule, not necessarily the principle of evidence, is used to determine whether there is evidence for or against ψ_* and this leads to the set $D(\psi_*) \subset \mathcal{X}$ as those data sets that do not give evidence in favor of $H_0 = \{\psi_*\}$. The rules of potential interest will satisfy $M(D(\psi_*) \mid \psi_*) \leq M(RB_\Psi(\psi_* \mid X) \leq 1 \mid \psi_*)$ since this implies better performance *a priori* in terms

of identifying when data has evidence in favor of H_0 via the set $D^c(\psi_*)$ than the principal of evidence. For example, $D(\psi_*) = \{x : RB_\Psi(\psi_* \,|\, x) \leq q\}$ for some $q < 1$ satisfies this but note that a value satisfying $q < RB_\Psi(\psi_* \,|\, x) \leq 1$ violates the principle of evidence if it is claimed there is evidence in favor of ψ_*. Putting $R(\psi_*) = \{x : RB_\Psi(\psi_* \,|\, x) \leq 1\}$ leads to the following result.

Theorem 1. (i) The prior probability $M(D(\psi_*))$ is maximized among all $D(\psi_*) \subset \mathcal{X}$ satisfying $M(D(\psi_*)\,|\,\psi_*) \leq M(R(\psi_*)\,|\,\psi_*)$ by $D(\psi_*) = R(\psi_*)$. (ii) If $\Pi_\Psi(\{\psi_*\}) = 0$, then $R(\psi_*)$ maximizes the prior probability of not obtaining evidence in favor of ψ_* when it is false and otherwise maximizes this probability among all rules satisfying $M(D(\psi_*)\,|\,\psi_*) = M(R(\psi_*)\,|\,\psi_*)$.

When $\Pi_\Psi(\{\psi_*\}) \neq 0$, rules may exist having greater prior probability of not getting evidence in favor of ψ_* when it is false but the price paid for this is the violation of the principle of evidence. Also, when comparing rules based on their ability to distinguish falsity it only seems fair that the rules perform the same under the truth. So Theorem 1 is a general optimality result for the principle of evidence applied to hypothesis assessment when considering bias against.

Now consider $C(x) = \{\psi : x \notin D(\psi)\}$ which is the set of ψ values for which there is evidence in their favor after observing x according to some alternative evidence rule. Since $M(\psi_* \notin C(X)\,|\,\psi) = M(D(\psi_*))\,|\,\psi)$, then $E_{\Pi_\Psi}\left(M(\psi \in C(X)\,|\,\psi)\right) = 1 - E_{\Pi_\Psi}\left(M(\psi \notin C(X)\,|\,\psi)\right) = 1 - E_{\Pi_\Psi}\left(M(D(\psi)\,|\,\psi)\right) \geq 1 - E_{\Pi_\Psi}\left(M(R(\psi)\,|\,\psi)\right) = E_{\Pi_\Psi}\left(M(\psi \in Pl_\Psi(X))\,|\,\psi)\right)$ and so the Bayesian coverage of C is at least as large as that of Pl_Ψ and so represents a viable alternative to using Pl_Ψ.

The following establishes an optimality result for Pl_Ψ.

Theorem 2. (i) The prior probability that the region C doesn't cover a value ψ_* generated from the prior, namely, $E_{\Pi_\Psi}(M(\psi_* \notin C(X)))$, is maximized among all regions satisfying $M(\psi_* \notin C(X)\,|\,\psi_*) \leq M(\psi_* \notin Pl_\Psi(X)\,|\,\psi_*)$ for every ψ_*, by $C = Pl_\Psi$. (ii) If $\Pi_\Psi(\{\psi_*\}) = 0$ for all ψ_*, then Pl_Ψ maximizes the prior probability of not covering a false value and otherwise maximizes this probability among all C satisfying $M(\psi_* \notin C(X)\,|\,\psi_*) = M(\psi_* \notin Pl_\Psi(X)\,|\,\psi_*)$ for all ψ_*.

Again when $\Pi_\Psi(\{\psi_*\}) \neq 0$ the existence of a region with better properties with respect to not covering false values than Pl_Ψ can't be ruled out but, when considering such a property, it seems only fair to compare regions with the same coverage probability and in that case Pl_Ψ is optimal. So Theorem 2 is also a general optimality result for the principle of evidence applied to estimation when considering bias against. Also, if there is a value $\psi_0 = \arg\inf_\psi M(\psi \in Pl_\Psi(X))\,|\,\psi)$, then $\gamma_0 = M(\psi_0 \in Pl_\Psi(X)\,|\,\psi_0)$ serves as a lower bound on the coverage probabilities, and thus Pl_Ψ is a γ_0-confidence region for ψ and this is a pure frequentist γ_0-confidence region when $\Psi(\theta) = \theta$. Since $M(\psi \in Pl_\Psi(X))\,|\,\psi) = 1 - M(\psi \notin Pl_\Psi(X))\,|\,\psi) = 1 - M(R(\psi_*)\,|\,\psi)$, then Example 1 shows that it is reasonable to expect that such a ψ_0 exists.

The principle of evidence leads to the following satisfying properties which connect the concept of bias as discussed here with the frequentist concept..

Theorem 3. (i) Using the principle of evidence, the prior probability of getting evidence in favor of ψ_* when it is true is greater than or equal to the prior

probability of getting evidence in favor of ψ_* given that ψ_* is false. (ii) The prior probability of Pl_Ψ covering the true value is always greater than or equal to the prior probability of Pl_Ψ covering a false value.

The properties stated in Theorem 3 are similar to a property called unbiasedness for frequentist procedures. For example, a test is unbiased if the probability of rejecting a null is always larger when it is false than when it is true and a confidence region is unbiased if the probability of covering the true value is always greater than the probability of covering a false value. While the inferences discussed here are "unbiased" in this generalized sense, they could still be biased against or in favor in the practical sense of this paper, as it is the amount of data that controls this.

Now consider bias in favor and suppose there is an alternative characterization of evidence that leads to the region $E(\psi_*)$ consisting of all data sets that do not lead to evidence against ψ_*. Putting $A(\psi_*) = \{x : RB_\Psi(\psi_* \,|\, x) \geq 1$, we restrict attention to regions satisfying $M(E(\psi_*) \,|\, \psi_*) \geq M(A(\psi_*) \,|\, \psi_*)$. Using (4) to measure bias in leads to the following results.

Theorem 4. (i) The prior probability $M(E(\psi_*))$ is minimized among all $E(\psi_*) \subset \mathcal{X}$ satisfying $M(E(\psi_*) \,|\, \psi_*) \geq M(A(\psi_*) \,|\, \psi_*)$ by $E(\psi_*) = A(\psi_*)$. (ii) If $\Pi_\Psi(\{\psi_*\}) = 0$, then the set $A(\psi_*)$ minimizes the prior probability of not obtaining evidence against ψ_* when it is false and otherwise minimizes this probability among all rules satisfying $M(E(\psi_*) \,|\, \psi_*) = M(A(\psi_*) \,|\, \psi_*)$.

Theorem 5. (i) The prior probability region C covers a value ψ_* generated from the prior, namely, $E_{\Pi_\Psi}(M(\psi_* \in C(X)))$, is minimized among all regions satisfying $M(\psi_* \in C(X) \,|\, \psi_*) \geq M(\psi_* \in Pl_\Psi(X) \,|\, \psi_*)$ for every ψ_*, by $C = Pl_\Psi$. (ii) If $\Pi_\Psi(\{\psi_*\}) = 0$ for all ψ_*, then Pl_Ψ minimizes the prior probability of covering a false value and otherwise minimizes this probability among all rules satisfying $M(\psi_* \in C(X) \,|\, \psi_*) = M(\psi_* \in Pl_\Psi(X) \,|\, \psi_*)$ for all ψ_*.

So Theorems 4 and 5 are optimality results for the principle of evidence when considering bias in favor.

Clearly the bias against H_0 is playing a role similar to size in frequentist statistics and the bias in favor is playing a role similar to power. A study that found evidence against H_0, but had a high bias against, or a study that found evidence in favor of H_0 but had a high bias in favor, could not be considered to be of high quality. Similarly, a study concerned with estimating a quantity of interest could not be considered of high quality if there is high bias against or in favor. There are some circumstances, however, where some bias is perhaps not an issue. For example, in a situation where sparsity is to be expected, then allowing for high bias in favor of certain hypotheses accompanied by low bias against, may be tolerable although this does reduce the reliability of any hypotheses where evidence is found in favor.

4 Examples

A number of examples are now considered.

Example 2. *Binomial.*

Suppose $x = (x_1, \ldots, x_n)$ is a sample from the Bernoulli(θ) with $\theta \in [0, 1]$ unknown so $n\bar{x} \sim$ binomial(n, θ) and interest is in θ. For the prior let $\theta \sim$ beta(α_0, β_0) where the hyperparameters are elicited as in, for example, Evans, Guttman and Li (2017), so $\theta \,|\, n\bar{x} \sim$ beta($\alpha_0 + n\bar{x}, \beta_0 + n(1 - \bar{x})$). Then

$$RB(\theta \,|\, n\bar{x}) = \frac{\Gamma(\alpha_0 + \beta_0 + n)}{\Gamma(\alpha_0 + n\bar{x})\Gamma(\beta_0 + n(1 - \bar{x}))} \frac{\Gamma(\alpha_0)\Gamma(\beta_0)}{\Gamma(\alpha_0 + \beta_0)} \theta^{n\bar{x}} (1 - \theta)^{n(1 - \bar{x})}$$

is unimodal with mode at $\bar{x}$, so $Pl(x)$ is an interval containing $\bar{x}$. Note that $M(\cdot \,|\, \theta)$ is the binomial(n, θ) probability measure and the bias against θ is given by $M(RB(\theta \,|\, n\bar{x}) \leq 1 \,|\, \theta)$ while the bias in favor of θ, using (5), is given by $\max M(RB(\theta \,|\, n\bar{x}) \geq 1 \,|\, \theta \pm \delta)$ for $\theta \in [\delta, 1 - \delta]$.

Consider first the prior given by $(\alpha_0, \beta_0) = (1, 1)$. Figure 4 gives the plots of the bias against for $n = 10$ (max. = 0.21, average = 0.11), $n = 50$ (max.= 0.07, average = 0.05) and $n = 100$ (max. = 0.05, average = 0.03). Therefore, when $n = 10$, then $Pl(x)$ is a 0.79-confidence interval for θ, when $n = 50$ it is a 0.93-confidence interval for θ and when $n = 100$ it is a 0.95-confidence interval for θ. For the informative prior given by $(\alpha_0, \beta_0) = (5, 5)$, Figure 5 gives the plots of the bias against for $n = 10$ (max. = 0.36, average = 0.21), $n = 50$ (max. = 0.16, average = 0.10) and $n = 100$ (max. = 0.11, average = 0.07). So when $n = 10$ then $Pl(x)$ is a 0.64-confidence interval for θ, when $n = 50$ it is a 0.84-confidence interval for θ and when $n = 100$ it is a 0.93-confidence interval for θ. One feature immediately stands out, namely, when using a more informative prior the bias against increases. As previously explained this phenomenon occurs because when the prior probability of θ is small, it is much easier to obtain evidence in favor than when the prior probability of θ is large.

Now consider bias in favor using (11). When $(\alpha_0, \beta_0) = (1, 1)$ and $\delta = 0.1$, Figure 6 gives the plots of the bias in favor with $\delta = 0.1$ for $n = 10$ (max. = 1.00, average = 0.84), $n = 50$ (max. = 0.72, average = 0.51) and $n = 100$ (max. = 0.50, average = 0.35). Therefore, when $n = 10$ the maximum probability that $Pl(x)$ contains a false value at least δ away from the true value is 1, when $n = 50$ this probability is 0.72 and when $n = 100$ it is a 0.50. When $(\alpha_0, \beta_0) = (5, 5)$ Figure 7 gives the plots of the bias in favor for $n = 10$ (max. = 1.00, average = 0.68), for $n = 50$ (max. = 1.00, average = 0.71) and for $n = 100$ (max. = 1.00, average = 0.49). So in this case the maximum probability that $Pl(x)$ contains a false value at least δ away from the true value is always 1, but when averaged with respect to the prior the values are considerably less. It is necessary to either increase n or δ to decrease bias in favor. For example, with $(\alpha_0, \beta_0) = (5, 5)$, $\delta = 0.1$ and $n = 400$ the maximum bias in favor is 0.02 and the average bias in favor is 0.02 and when $n = 600$ these quantities equal 0 to two decimals. When $\delta = 0.2$ and $n = 50$ the maximum bias in favor is 0.29 and the average bias in favor is 0.11 and when $n = 100$ the maximum bias in favor is 0.01 and the average bias in favor is 0.01.

Example 3. *Location-scale normal - quantiles.*

Suppose $x = (x_1, \ldots, x_n)$ is a sample from $N(\mu, \sigma^2)$ with $(\mu, \sigma^2) \in R^1 \times (0, \infty)$ unknown with prior $\mu \,|\, \sigma^2 \sim N(\mu_0, \tau_0^2 \sigma^2), \sigma^{-2} \sim$ gamma$_{\text{rate}}(\alpha_0, \beta_0)$. The
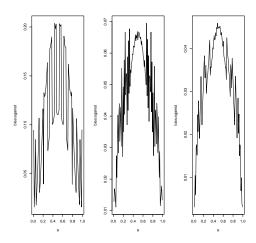
Figure 4: Plots of bias against as a function of θ for $n = 10, 50$ and 100 when using beta$(1, 1)$ prior.
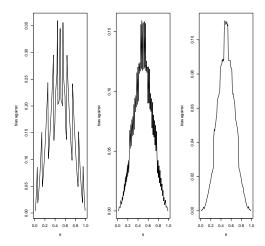

Figure 5: Plots of bias against as a function of θ for $n = 10, 50$ and 100 when using beta$(5, 5)$ prior.
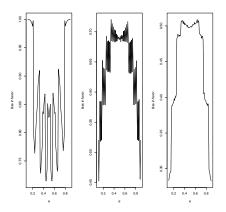
17

Figure 6: Plots of bias in favor as a function of θ for $n = 10, 50$ and 100 when using beta$(1, 1)$ prior with $\delta = 0.1$.
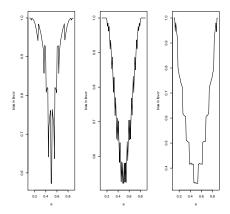

Figure 7: Plots of bias in favor as a function of θ for $n = 10, 50$ and 100 when using beta$(5, 5)$ prior with $\delta = 0.1$.

hyperparameters $(\mu_0, \tau_0^2, \alpha_0, \beta_0)$ can be obtained via an elicitation as, for example, discussed in Evans and Tomal (2018) for the more general regression model. This example is easily generalized to the regression context. A MSS is $T(x) = (\bar{x}, ||x - \bar{x}1||^2)$ with the posterior distribution given by $\mu \,|\, \sigma^2, T(x) \sim N(\mu_{0x}, (n + 1/\tau_0^2)^{-1} \sigma^2), \sigma^{-2} \,|\, T(x) \sim \text{gamma}_{rate}(\alpha_0 + n/2, \beta_{0x})$ where $\mu_{0x} = (n + 1/\tau_0^2)^{-1}(n\bar{x} + \mu_0/\tau_0^2)$ and $\beta_{0x} = \beta_0 + ||x - \bar{x}1||^2/2 + n(\bar{x} - \mu_0)^2/2(n\tau_0^2 + 1)$.

Suppose interest is in the γ-th quantile $\psi = \Psi(\mu, \sigma^2) = \mu + \sigma z_\gamma$, where $z_\gamma = \Phi^{-1}(\gamma)$. To determine the bias for or against ψ we need the prior and posterior of ψ which in this case cannot be worked out in closed form. It is easy, however, to work with the discretized ψ by simply generating from the prior and posterior of (μ, σ^2), estimate the contents of the relevant intervals and then approximate the relative belief ratio using these. A natural approach to the discretization is to base it on the prior mean $E(\psi) = \mu_0 + \beta_0^{1/2}(\Gamma(\alpha_0 - 1/2)/\Gamma(\alpha_0))z_\gamma$ and variance $Var(\psi) = E(\psi^2) - (E(\psi))^2$ where $E(\psi^2) = (z_\gamma^2 + \tau_0^2)\beta_0/(\alpha_0 - 1)$. So for a given δ, we discretize using $2k + 1$ intervals $(E(\psi) + i\delta, E(\psi) + (i + 1)\delta]$ where $k = cSD(\psi)/\delta$ and c is chosen so that the collection of intervals covers the effective support of ψ which is easily assessed as part of the simulation. For example, with the prior given by hyperparameters $\mu_0 = 0, \tau_0^2 = 1, \alpha_0 = 2, \beta_0 = 1$ and $\gamma = 0.5, \delta = 0.1, c = 5$, then $k = 50$ and, on generating 10^5 values from the prior, these intervals contained $99,699$ of the values and with $c = 6$, then $k = 60$ and these intervals contained $99,901$ of the generated values. Similar results are obtained for more extreme quantiles and this is because the intervals shift with the quantile.

For the bias against for estimation the value of $M(RB_\Psi(\psi \,|\, X) \leq 1 \,|\, \psi)$ is needed for a range of ψ values. For this we need to generate from the conditional prior distribution of T given $\Psi(\mu, \sigma^2) = \psi$ and an algorithm for generating from the conditional prior of (μ, σ^2) given ψ is needed. Putting $\nu = 1/\sigma^2$, the transformation $(\mu, \nu) \to (\psi, \nu) = (\mu + \nu^{-1/2}z_\gamma, \nu)$ has Jacobian equal to 1, so the conditional prior distribution of $\nu \,|\, \psi$ has density proportional to $\nu^{\alpha_0 - 1/2} \exp\{-\beta_0\nu\} \exp\{-\nu\left(\psi - \mu_0 - \nu^{-1/2}z_\gamma\right)^2/2\tau_0^2\}$. The following gives a rejection algorithm for generating from this distribution:

1. generate $\nu \sim \text{gamma}(\alpha_0 + 1/2, \beta_0)$, 2. generate $u \sim \text{unif}(0, 1)$ independent of ν, 3. if $u \leq \exp\{-\nu\left(\psi - \mu_0 - \nu^{-1/2}z_\gamma\right)^2/2\tau_0^2\}$ return ν, else go to 1.

As ψ moves away from the prior expected value $E(\psi)$ this algorithm becomes less efficient but even when the expected number of iterations is 86 (when $\gamma = 0.95, \psi = 12$), generating a sample of 10^4 is almost instantaneous. Figure 8 is a plot of the conditional prior of ν given that $\psi = 2$. After generating ν then generate $||x - \bar{x}1||^2 \sim \nu^{-1}\text{chi-squared}(n - 1)$ and $\bar{x} \sim N(\psi - \nu^{-1/2}z_\gamma, \nu^{-1}/n)$ to complete the generation of a value from $M_T(\cdot \,|\, \psi)$.

The bias against as a function of $\psi = \mu + \sigma z_{0.95}$, has maximum value 0.151 when $n = 10$ and so $Pl_\Psi(x)$ is a 0.849-confidence region for ψ while the average bias against is 0.104 so the Bayesian coverage is 0.896. Table 6 gives the coverages for other values of n as well. Figure 9 is a plot of the bias in favor as a function of ψ with $\delta = \pm 0.5$ and $n = 10$. The average bias in favor is 0.629304. When
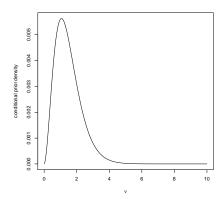
Figure 8: Conditional prior density of $\nu = 1/\sigma^2$ given $\psi = 2$ when $\gamma = 0.95$ and $\mu_0 = 0, \tau_0^2 = 1, \alpha_0 = 2, \beta_0 = 1$.

n	Frequentist coverage	Bayesian coverage
10	0.849	0.896
20	0.895	0.927
50	0.934	0.958
100	0.955	0.973

Table 6: Coverage probabilities for $Pl_\psi(x)$ for the 0.95 quantile in Example 2.

$n = 50$ the average bias in favor is 0.3348178.

The case $\gamma = 0.50$, so $\psi = \Psi(\mu, \sigma^2) = \mu$, is also of interest. For $n = 10$ then $Pl_\Psi(x)$ has 0.878 frequentist coverage and 0.926 Bayesian coverage, when $n = 20$ the coverages are 0.916 and 0.952 while when $n = 50$ the coverages are 0.950 and 0.973. When $n = 10, \delta = 0.5$ the average bias in favor is 0.619, when $n = 20$ this is 0.4206 and for $n = 100$ the average bias in favor is 0.091.

Example 4. *Normal Regression - prediction.*

Prediction problems have some unique aspects when compared to inferences about parameters. To see this consider first the location normal model of Example 1 and suppose the problem is to make inference about a future value $y \sim N(\mu, \sigma_0^2)$. The prior predictive distribution is $y \sim N(\mu_0, \tau_0^2 + \sigma_0^2)$ and the posterior predictive is $y \sim N(\mu_x, \sigma_n^2 + \sigma_0^2)$ where $\mu_x = \sigma_n^2(n\bar{x}/\sigma_0^2 + \mu_0/\tau_0^2), \sigma_n^2 = \left(n/\sigma_0^2 + 1/\tau_0^2\right)^{-1}$ and so

$$RB(y \,|\, \bar{x}) = \left(\frac{\tau_0^2 + \sigma_0^2}{\sigma_n^2 + \sigma_0^2}\right)^{1/2} \exp\left\{-\frac{1}{2}\left[\frac{(y - \mu_x)^2}{\sigma_n^2 + \sigma_0^2} - \frac{(y - \mu_0)^2}{\tau_0^2 + \sigma_0^2}\right]\right\}.$$

For a given y the bias against is given by $M(RB(y \,|\, \bar{x}) \leq 1 \,|\, y)$ and for this we need the conditional prior predictive of $\bar{x} \,|\, y$. The joint prior predictive is
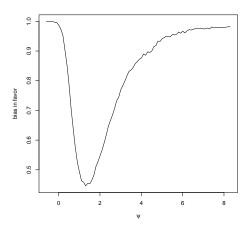
Figure 9: The bias in favor as a function of ψ when $n = 10, \delta = 0.5$ and using a prior with hyperparameters $\mu_0 = 0, \tau_0^2 = 1, \alpha_0 = 2, \beta_0 = 1$.

$(\bar{x}, y) \sim N_2(\mu_0 1_2, \Sigma_0)$ where

$$\Sigma_0 = \left(\begin{array}{cc} \tau_0^2 + \sigma_0^2/n & \tau_0^2 \\ \tau_0^2 & \tau_0^2 + \sigma_0^2 \end{array} \right)$$

and so $\bar{x} \mid y \sim N(\mu_0 + \tau_0^2(y - \mu_0)/(\tau_0^2 + \sigma_0^2), \sigma_0^2 \left(\tau_0^2/(\tau_0^2 + \sigma_0^2) + 1/n \right))$. From this we see that, as $n \to \infty$ the conditional prior distribution of $\mu_x \mid y$ converges to the $N \left(\mu_0 + \tau_0^2(y - \mu_0)/(\tau_0^2 + \sigma_0^2), \sigma_0^2 \tau_0^2/(\tau_0^2 + \sigma_0^2) \right)$ distribution. Then with $Z \sim N(0, 1)$ and $r = \tau_0^2/\sigma_0^2$, putting $d((y - \mu_0)/\sigma_0, r) = (1 + 1/r) \log (1 + r) + r^{-1}(y - \mu_0)^2/\sigma_0^2$

$$M(RB(y \mid \bar{x}) \leq 1 \mid y) \to 1 - P \left(Z \in \left[\begin{array}{c} r^{-1/2} (1 + r)^{-1/2} \left(\frac{y - \mu_0}{\sigma_0} \right) \pm \\ d^{1/2} \left(\frac{y - \mu_0}{\sigma_0}, r \right) \end{array} \right] \right)$$

as $n \to \infty$. So the bias against does not go to 0 as $n \to \infty$ and there is a limiting lower bound to the prior probability that evidence in favor of a specific y will not be obtained. This baseline is dependent on both $(y - \mu_0)/\sigma_0$ and r. As $r = \tau_0^2/\sigma_0^2 \to \infty$ this baseline bias against goes to 0 and so it is necessary to ensure that the prior variance is not too small. Table (7) gives some values for the bias against and it is seen that if τ_0^2/σ_0^2 is too small, then there is substantial bias against even when y is a reasonable value from the distribution. When $\tau_0^2/\sigma_0^2 = 1, (y - \mu_0)/\sigma_0 = 0$ and $n = 10$ the bias against is computed to be 0.248 which is quite close to the baseline so increasing sample size will not reduce bias against by much and similar results are obtained for the other cases.

Now consider bias in favor of y, namely, $M(RB(y \mid \bar{x}) \geq 1 \mid y \pm \delta)$ for some choice of δ. False values for y correspond to values in the tails so we consider,

21

τ_0^2/σ_0^2	bias against $\frac{y-\mu_0}{\sigma_0}=0$	bias against $\frac{y-\mu_0}{\sigma_0}=1$
1	0.239	0.213
10	0.104	0.100
100	0.031	0.031
1/2	0.270	0.263
1/100	0.316	0.460

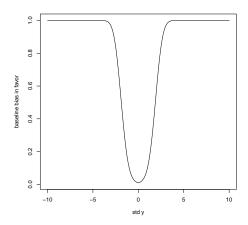Table 7: Baseline bias against values for prediiction for location normal in Example 4.

Figure 10: Plot of the baseline bias in favor for values of $(y-\mu_0)/\sigma_0$ when $\tau_0^2/\sigma_0^2 = 1$ when $\delta = 5$.

for example, $y + \delta$ as a value in the central region of the prior and then a large value of δ puts y in the tails. Again the bias in favor has a baseline value as $n \to \infty$. A similar argument leads to the bias in favor of y satisfying

$$M(RB(y\,|\,\bar{x}) \geq 1\,|\,y \pm \delta)$$
$$\to P\left(Z \in \left[r^{-1/2}\,(1+r)^{-1/2}\left(\frac{y-\mu_0}{\sigma_0} \pm r\frac{\delta}{\sigma_0}\right) \pm d^{1/2}\left(\frac{y-\mu_0}{\sigma_0}, r\right)\right]\right).$$

Figure 10 is a plot of $\sup M(RB(y\,|\,\bar{x}) \geq 1\,|\,y \pm \delta)$. So the bias in favor is reasonably low for central values of y but it is to be noted that once again there is a trade-off as when τ increases the bias in favor goes to 1.

Prediction plays a bigger role in regression problems but we can expect the same issues to apply as in the location problem. Suppose $y \sim N_n(X\beta, \sigma^2 I)$ where $X \in R^{n \times k}$ is of rank k, $(\beta, \sigma^2) \in R^k \times (0, \infty)$ is unknown, our interest is in predicting a future value $y_{new} \sim N(w^t\beta, \sigma^2)$ for some fixed known w and, putting $\nu = 1/\sigma^2$, the conjugate prior $\beta\,|\,\nu \sim N_k(\beta_0, \nu^{-1}\Sigma_0)$, $\nu \sim$

22

gamma$_{\text{rate}}(\alpha_0, \eta_0)$ is used. Specifying the hyperparameters $(\beta_0, \Sigma_0, \alpha_0, \eta_0)$ can be carried out using an elicitation algorithm such as that discussed in Evans and Tomal (2018).

For the bias calculations it is necessary to generate values of the MSS $(b, s^2) = ((X^t X)^{-1} X^t y, ||y - Xb||^2)$ from the conditional prior predictive $M(\cdot \,|\, y_{new})$. This is accomplished by generating from the conditional prior of $(\beta, \nu) \,|\, y_{new}$ and then generating $b \sim N_k(\beta, \nu^{-1}(X^t X)^{-1})$ independent of $s^2 \sim \nu^{-1}$ chi-squared$(n - k)$. The conditional prior of $(\beta, \nu) \,|\, y_{new}$ is proportional to

$$\nu^{\alpha_0 - 1/2} \exp\{-\eta_0(y_{new})\nu\} \times$$
$$\nu^{k/2} \exp\left\{-\frac{\nu}{2} \left(\beta - \left(\Sigma_0^{-1} + ww^t\right)^{-1} (\Sigma_0^{-1}\beta_0 + y_{new}w)\right)^t \left(\Sigma_0^{-1} + ww^t\right) (\cdot)\right\}$$

where, using $\left(\Sigma_0^{-1} + ww^t\right)^{-1} = \Sigma_0 - (1 + w^t\Sigma_0 w)^{-1}\Sigma_0 ww^t\Sigma_0$, $\eta_0(y_{new}) = \eta_0 + (1 + w^t\Sigma_0 w)^{-1}(w^t\beta - y_{new})^2/2$. So generating $(\beta, \nu) \,|\, y_{new}$ is accomplished via $\nu \sim$ gamma$_{\text{rate}}(\alpha_0 + 1/2, \eta_0(y_{new}))$,

$$\beta \,|\, \nu \sim N_k\left(\left(I - \frac{\Sigma_0 ww^t}{1 + w^t\Sigma_0 w}\right)(\beta_0 + y_{new}\Sigma_0 w), \nu^{-1}\left(\Sigma_0 - \frac{\Sigma_0 ww^t\Sigma_0}{1 + w^t\Sigma_0 w}\right)\right).$$

For each generated (b, s^2) it is necessary to compute the relative belief ratio $RB(y_{new} \,|\, b, s^2)$ and determine if it is less than or equal to 1. There are closed forms for the prior and conditional densities of y_{new} since

$$y_{new} \sim w^t\beta_0 + \left\{\eta_0(1 + w^t\Sigma_0 w)/\alpha_0\right\}^{1/2} t_{2\alpha_0},$$
$$y_{new} \,|\, (b, s^2) \sim w^t\beta_0(b, s^2) + \left\{\frac{\eta_0(b, s^2)(1 + w^t\left(\Sigma_0^{-1} + X^t X\right)^{-1} w)}{\alpha_0 + n/2}\right\}^{1/2} t_{2\alpha_0 + n}$$

where t_λ denotes a Student(λ) random variable and

$$\beta_0(b, s^2) = \left(\Sigma_0^{-1} + X^t X\right)^{-1} \left(\Sigma_0^{-1}\beta_0 + X^t Xb\right)$$
$$\eta_0(b, s^2) = \eta_0 + \left[s^2 + ||Xb||^2 + ||\Sigma_0^{-1}\beta_0||^2 - \beta_0(b, s^2)^t \left(\Sigma_0^{-1} + X^t X\right) \beta_0(b, s^2)\right]/2.$$

These results permit the calculation of the biases as in the location problem.

5   Conclusions

There are several conclusions that can be drawn from the discussion here. First, it is necessary to take bias into account when considering Bayesian procedures and currently this is generally not being done. Depending on the purpose of the study, some values concerning both bias against and bias in favor need to be quoted as these are figures of merit for the study. The approach to Bayesian inferences via a characterization of evidence makes this relatively straight-forward conceptually. Second, frequentism plays a role in Bayesian statistical reasoning, not through the inferences, but rather through the design as it how we

determine and control the biases. Overall this makes sense because, before the data is seen, it is natural to be concerned about what inferences can be reliably drawn. Once the data is observed, however, it is the evidence in this data set that matters and not the evidence in the data sets not seen. Still, if we ignore the latter it may be that the existence of bias makes the inferences drawn of very low quality. Third, the results concerning the standard p-value in Example 1 can be seen to apply quite generally and this makes any discussion about how to characterize and measure evidence of considerable importance. The principle of evidence makes a substantial contribution in this regard as was shown in a variety of results. The major purpose of this paper, however, is to deal with a key criticism of Bayesian methodology, namely, that inferences can be biased because of their dependence on the subjective beliefs of the analyst. This criticism is accepted, but we also assert that this can be dealt with in a logical and scientific fashion as has been demonstrated in this paper.

6   References

Baskurt, Z. and Evans, M. (2013) Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. Bayesian Analysis, 8, 3, 569-590.

Berger, J.O. and Selke, T. (1987) Testing a point null hypothesis: the irreconcilability of p values and evidence. Journal of the American Statistical Association, 82, 397, 112-122.

Berger, J.O. and Delampady, M. (1987) Testing precise hypotheses. Statistical Science, 2, 3, 317-335.

Cousins, R.D. (2017) The Jeffreys–Lindley paradox and discovery criteria in high energy physics. Synthese, 194, 2, 395–432.

Evans, M. (2015) Measuring Statistical Evidence Using Relative Belief. Monographs on Statistics and Applied Probability 144, CRC Press.

Evans, M., Guttman, I. and Li, P. (2017) Prior elicitation, assessment and inference with a Dirichlet prior. Entropy 2017, 19(10), 564; doi:10.3390/e1910056.

Evans, M. and Tomal, J. (2018) Multiple testing via relative belief ratios. FACETS, 3: 563-583, DOI: 10.1139/facets-2017-0121.

Gu, Y., Li, W. Evans, M. and Englert, B-G. (2019) Very strong evidence in favor of quantum mechanics and against local hidden variables from a Bayesian analysis. Physical Review A 99, 022112(1-17).

Robert, C. P. (2014) On the Jeffreys-Lindley paradox. Philosophy of Science, 81, 216–232.

Shafer, G. (1982) Lindley's paradox (with discussion). Journal of the American Statistical Association, 77, 378, 325-351.

Spanos, A. Who should be afraid of the Jeffreys-Lindley paradox? Philosophy of Science, 80, 1, 73 - 93.

Sprenger, J. (2013) Testing a precise null hypothesis: The case of Lindley's paradox. Philosophy of Science, 80, 733-744.

Villa, C. and Walker, S. (2017) On the mathematics of the Jeffreys–Lindley

paradox. Communications in Statistics - Theory and Methods, 46, 24, 12290-12298.

7 Appendix

Proof of Theorem 1. The Savage-Dickey ratio result implies $RB_\Psi(\psi_* \,|\, x) = m_{\psi_*}(x)/m(x)$ and note $R(\psi_*) = \{x : m_{\psi_*}(x) \le m(x)\}$. Now $\mathcal{X}_1 = \{x : I_{R(\psi_*)}(x) - I_{D(\psi_*)}(x) < 0\} = \{x : I_{R(\psi_*)} - I_{D(\psi_*)}(x) < 0, m_{\psi_*}(x) > m(x)\}$ and $\mathcal{X}_2 = \{x : I_{R(\psi_*)}(x) - I_{D(\psi_*)}(x) > 0\} = \{x : I_{R(\psi_*)}(x) - I_{D(\psi_*)}(x) \ge 0, m_{\psi_*}(x) \le m(x)\}$. Then $M(R(\psi_*)) - M(D(\psi_*)) = \int_{\mathcal{X}_1}(I_{R(\psi_*)}(x) - I_{D(\psi_*)}(x)) M(dx) + \int_{\mathcal{X}_2}(I_{R(\psi_*)}(x) - I_{D(\psi_*)}(x)) M(dx) \ge M(R(\psi_*) \,|\, \psi_*) - M(D(\psi_*) \,|\, \psi_*) \ge 0$ establishing (i). Also, $M(D(\psi_*)) = M(D(\psi_*) \,|\, \psi_*)\Pi_\Psi(\{\psi_*\}) + \int_{\Psi \setminus \{\psi_*\}} M(D(\psi_*) \,|\, \psi)\,\Pi_\Psi(d\psi)$ and the integral is the prior probability of not getting evidence in favor of ψ_* when it is false and this establishes (ii).

Proof of Theorem 2. Now $E_{\Pi_\Psi}(M(\psi_* \notin C(X))) = E_{\Pi_\Psi^2}(M(\psi_* \notin C(X) \,|\, \psi)) = E_{\Pi_\Psi^2}(M(D(\psi_*)) \,|\, \psi)) = \int_\Psi M(D(\psi_*))\,\Pi_\Psi(d\psi_*)$ and (i) follows from Theorem 1. Also, $\int_\Psi M(D(\psi_*))\,\Pi_\Psi(d\psi_*) = E_{\Pi_\Psi}(\int_\Psi M(D(\psi_*) \,|\, \psi)\,\Pi_\Psi(d\psi)) = E_{\Pi_\Psi}(M(D(\psi_*) \,|\, \psi_*)\Pi_\Psi(\{\psi_*\})) + E_{\Pi_\Psi}(\int_{\Psi \setminus \{\psi_*\}} M(D(\psi_*) \,|\, \psi)\,\Pi_\Psi(d\psi)) = E_{\Pi_\Psi}(M(\psi_* \notin C(X) \,|\, \psi_*)\Pi_\Psi(\{\psi_*\})) + E_{\Pi_\Psi}(\int_{\Psi \setminus \{\psi_*\}} M(\psi_* \notin C(X) \,|\, \psi)\,\Pi_\Psi(d\psi))$ establishing (ii).

Proof of Theorem 3. Now $M(R(\psi_*) \,|\, \psi_*) = \int I_{R(\psi_*)}(x)\,M_{\psi_*}(dx) \le \int I_{R(\psi_*)}(x)\,M(dx) = M(R(\psi_*)) = \int_\Psi M(R(\psi_*) \,|\, \psi)\,\Pi(d\psi) = M(R(\psi_*) \,|\, \psi_*)\Pi_\Psi(\{\psi_*\}) + \int_{\Psi \setminus \{\psi_*\}} M(R(\psi_*) \,|\, \psi)\,\Pi_\Psi(d\psi)$ and so $\Pi_\Psi(\{\psi_*\}^c)M(R(\psi_*) \,|\, \psi_*) \le \int_{\Psi \setminus \{\psi_*\}} M(R(\psi_*) \,|\, \psi)\,\Pi_\Psi(d\psi)$ which implies (i). Furthermore, (ii) is implied by

$$E_{\Pi_\Psi}(M(\psi_* \notin Pl_\Psi(X) \,|\, \psi_*)) = E_{\Pi_\Psi}(M(R(\psi_*) \,|\, \psi_*))$$

$$\le E_{\Pi_\Psi}\left(\int_{\Psi \setminus \{\psi_*\}} M(R(\psi_*) \,|\, \psi)\,\Pi_\Psi(d\psi)/\Pi_\Psi(\{\psi_*\}^c)\right)$$

$$= E_{\Pi_\Psi}\left(\int_{\Psi \setminus \{\psi_*\}} M(\psi_* \notin Pl_\Psi(X) \,|\, \psi)\,\Pi_\Psi(d\psi)/\Pi_\Psi(\{\psi_*\}^c)\right).$$

Proof of Theorem 4. It is easy to see that the proof of Theorem 1 can be modified to show that among all regions $D^{int}(\psi_*) \subset \mathcal{X}$ satisfying $M(D^{int}(\psi_*) \,|\, \psi_*) \le M(RB_\Psi(\psi_* \,|\, X) < 1 \,|\, \psi_*)$ the prior probability $M(D^{int}(\psi_*))$ is maximized by $D^{int}(\psi_*) = \{x : RB_\Psi(\psi_* \,|\, x) < 1\}$. This clearly implies (i) and (ii) follows similarly.

Proof of Theorem 5. Now $E_{\Pi_\Psi}(M(\psi_* \in C(X))) = E_{\Pi_\Psi^2}(M(\psi_* \in C(X) \,|\, \psi)) = E_{\Pi_\Psi^2}(M(D^c(\psi_*)) \,|\, \psi)) = E_{\Pi_\Psi}(M(D^c(\psi_*))$ and (i) follows from Theorem 1(i). Also, (ii) is implied by $E_{\Pi_\Psi}(M(D^c(\psi_*)) = \int_\Psi M(D^c(\psi_*) \,|\, \psi_*)\Pi_\Psi(\{\psi_*\})\,\Pi_\Psi(d\psi_*) + \int_\Psi \int_{\Psi \setminus \{\psi_*\}} M(D^c(\psi_*) \,|\, \psi)\,\Pi_\Psi(d\psi)\,\Pi_\Psi(d\psi_*) = \int_\Psi M(\psi_* \in C(X) \,|\, \psi_*)\Pi_\Psi(\{\psi_*\})\,\Pi_\Psi(d\psi_*) + \int_\Psi \int_{\Psi \setminus \{\psi_*\}} M(\psi_* \in C(X) \,|\, \psi)\,\Pi_\Psi(d\psi)\,\Pi_\Psi(d\psi_*).$