Neutral-data comparisons defining a spectrum between Bayes factors and the Schwarz criterion

Dan J. Spitzner

April 7, 2014

Abstract

In testing problems, neutral-data comparisons assess evidence in a parallel manner as Bayes factors, but are drastically less sensitive to scale parameters of the prior, and are thus suitable for use with vague priors. This article proposes a calibration rule for neutral-data comparisons that is motivated from a well known connection between unit-information priors and the Schwarz criterion. These ideas are examined and illustrated on data exemplifying the Behrens-Fisher problem and in the analysis of two-way tables.

KEY WORDS: Bayesian testing; Bayes factors; neutral-data comparisons; unit-information priors; Schwarz criterion.

ABBREVIATED TITLE: Neutral-data comparisons defining a spectrum.

1 Introduction

Neutral-data comparisons were introduced in Spitzner (2011) as an approach to Bayesian testing that avoids concerns over the sensitivity of Bayes factors to prior dispersion. These quantities are, for present purposes, *calibrated* Bayes factors (although they have an independent interpretation), whose calibration targets an imaginary "neutral" data set that is presumed to exhibit no more evidence for one model under test than the other. If $BF_{01}(\boldsymbol{Y})$ is the Bayes factor assessing models $\mathcal{M}_0$ *vs* $\mathcal{M}_1$, the corresponding neutral-data comparison is

$$NDC_{01}(\boldsymbol{Y}) = BF_{01}(\boldsymbol{Y})/BF_{01}(\tilde{\boldsymbol{Y}}), \tag{1}$$

where $\tilde{\boldsymbol{Y}}$ is neutral data. See Spitzner (2011) for additional perspectives on neutral data and $NDC_{01}(\boldsymbol{Y})$. A short summary of the underlying motivation for (1) is provided in the appendix.

A central concern in the use of neutral-data comparisons is the manner of specifying $\tilde{\boldsymbol{Y}}$. This article explores a proposed scheme that connects neutral-data comparisons to the criterion of Schwarz (1978) (a.k.a, the Bayesian information criterion, BIC), as it is understood through Kass and Wasserman's (1995) theory of "unit-information priors." To illustrate, consider the simple Gaussian case in which the data are $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, for independent Y_i, the model $\mathcal{M}_0$ has $Y_i \sim G(0,1)$, and model $\mathcal{M}_1$ has $Y_i|\theta \sim G(\theta, 1)$ with $\theta \sim G(0, \tau^2)$. The Bayes factor in this case is

$$BF_{01}(\boldsymbol{Y}) = (1+\tau^2 n)^{1/2} e^{-\frac{1}{2} w Z^2}, \tag{2}$$

where $w = \tau^2 n/(1+\tau^2 n)$ and $Z = n^{-1/2}\sum_{i=1}^n Y_i$. Kass and Wasserman (1995) highlight the particular setting $\tau^2 = 1$ as defining a "unit-information prior" within this context, and note that the log-Bayes factor is then approximated by the corresponding Schwarz criterion statistic,

$$S_{01}(\boldsymbol{Y}) = -\frac{1}{2} Z^2 + \frac{1}{2}\log n \tag{3}$$

Dan Spitzner is Associate Professor, Department of Statistics, University of Virginia, P. O. Box 400135, Charlottesville, VA 22904-4135, USA. The author is grateful for invaluable support in preparing this technical report from the National Science Foundation (grant number SES-1260803).

Noting certain broad generalizations of this phenomenon, Kass and Wasserman conclude that $\exp S_{01}(\boldsymbol{Y})$ is an "interesting approximate Bayes factor, and thus a potentially useful quantification of evidence." (p. 928). Upon setting $\tilde{\boldsymbol{Y}}$ in such a way that the neutral-data version of Z^2 is $\tilde{Z}^2 = w^{-1}\log(1/\tau^2 + n)$, the resulting neutral-data comparison is

$$NDC_{01}(\boldsymbol{Y}) = (1/\tau^2 + n)^{1/2} e^{-\frac{1}{2}wZ^2}. \tag{4}$$

One sees from this formula that $NDC_{01}(\boldsymbol{Y}) = BF_{01}(\boldsymbol{Y})$ when $\tau^2 = 1$, and $NDC_{01}(\boldsymbol{Y}) \to \exp S_{01}(\boldsymbol{Y})$ as $\tau^2 \to \infty$. Thus, a neutral-data comparison, in its formulation above, defines a spectrum of assessments, indexed by τ^2, whose endpoint on one side is a Bayes factor and the exponentiated Schwarz criterion on the other.

Such positioning highlights how a neutral-data comparison modulates the sensitivity of a Bayes factor to prior dispersion. Whereas drastic changes in the prior scale parameter, τ^2, will result in similarly drastic changes in the value produced by a Bayes factor, the range of possible values produced by a neutral-data comparison roughly matches the variation typically observed in Bayesian estimation, *e.g.*, in the posterior mean under $\mathcal{M}_1$. Relative to the Schwarz criterion, the advantage of neutral-data comparisons is that it depends on τ^2, and is therefore responsive to subjective knowledge, as would be desired of any Bayesian procedure. The methodology developed in this article is therefore intended for scenarios in which some crude picture of prior knowledge is available, but a precise articulation has not yet been made, either due to lack of resources or lack of access to a suitable expert. (The author conjectures that this is the most widespread scenario encountered in statistical practice.) Reflecting the two endpoints of the spectrum alluded to above, the statistic (1) would be justly labelled a "robust" Bayes factor or an exponentiated "subjective" Schwartz criterion.

Closer examination of the neutral-data comparison (4) hints at the central idea of the proposed scheme. By solving (1), the reader quickly sees that the imaginary data $\tilde{\boldsymbol{Y}}$ has been chosen to make $BF_{01}(\tilde{\boldsymbol{Y}}) = \tau$, which is relevant for suggesting a certain precise characterization of "neutrality." In particular, one should notice that $BF_{01}(\tilde{\boldsymbol{Y}}) = 1$ precisely when $\tau^2 = 1$, the setting of a unit-information prior. The proposed calibration rule is thereby revealed to be that *neutral data are to yield neutral evidence under a unit-information prior*. The task carried out in the discussion below to extend this rule to broader scenarios, including those involving non-Gaussian likelihood functions and nuisance parameters.

The reader may ponder a straightforward modification of the calibration rule as follows: "neutral data are to yield neutral evidence under a X prior," where X is any prior that is well established as serving in the role of a default prior for Bayes factors. For instance, the priors associated with the intrinsic Bayes factor of Berger and Pericchi (1996) and fractional Bayes factor of O'Hagan (1995) are notable potential substitutes for X, and will be examined briefly in what follows. Nevertheless, unit-information priors are appealing for present purposes due to their connection to the Schwarz criterion. The unit-scale Cauchy prior, a default prior proposed by Jeffreys (1961), may be regarded as a non-Gaussian unit-information prior, and so will be covered in the framework developed below. Other related literature include Spiegelhalter and Smith's (1982) discussion of Bayes factors derived from improper priors, whose ideas are adapted in Spitzner (2014) to neutral-data comparisons for variable selection. Additional discussion of the Schwarz criterion, especially of its use in practice, is found in Raftery (1995) and Weakliem (1999); recent developments appear in Bollen *et al.* (2012). See also Lu (2012), in which unit-information priors are used in an interesting way to rethink intrinsic Bayes factors. The use of imaginary data is often credited to Good (1950), who calls it the "device of imaginary results."

In what follows, Section 2 presents the article's main formulas, including a broad criterion for selecting neutral data that is suitable for testing problems that involve regular likelihood functions. Section 3 follows with illustrations of the proposed data-analysis methodology to the classical Behrens-Fisher problem and to the analysis of count data in a two-way table. Concluding discussion appears in Section 4.

2 Main results

The article's approach to handling nuisance parameters aligns more closely to the applied framework of Bayes factors rather than that of the Schwartz criterion. In particular, nuisance parameters are not to be replaced by "plug-in" estimates, as is sometimes done in applications of the Schwarz criterion, but are regarded as quantities to be conditioned upon during analysis formulation, and integrated across when calculating analysis results. Thus, the starting point in developing the proposed methodology are conditional

versions of the Bayes factor and neutral-data comparison,

$$BF_{01}(\boldsymbol{Y}|\phi) = \frac{\pi_0(\boldsymbol{Y}|\phi)}{\pi_1(\boldsymbol{Y}|\phi)} \quad \text{and} \quad NDC_{01}(\boldsymbol{Y}|\phi) = \frac{BF_{01}(\boldsymbol{Y}|\phi)}{BF_{01}(\tilde{\boldsymbol{Y}}|\phi)}, \tag{5}$$

in which ϕ is a nuisance parameter and $\pi_i(\boldsymbol{Y}|\phi)$ is a marginal density for the data under model $\mathcal{M}_i$, conditional on ϕ. In this formulation, neutral data are also specified conditionally on the nuisance parameter, $\tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{Y}}(\phi)$, but that dependency is omitted in the notation. Denote by $\boldsymbol{\theta}$ the parameter under test in assessing $\mathcal{M}_0$ *vs* $\mathcal{M}_1$, for which the "null" setting $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ is associated with the model $\mathcal{M}_0$. The log-likelihood function is $l(\boldsymbol{\theta}, \phi; \boldsymbol{Y})$, hence the marginal data-densities are $\pi_0(\boldsymbol{Y}|\phi) = l(\boldsymbol{\theta}_0, \phi; \boldsymbol{Y})$ and $\pi_1(\boldsymbol{Y}|\phi) = \int l(\boldsymbol{\theta}, \phi; \boldsymbol{Y})\pi(\boldsymbol{\theta}|\phi)d\boldsymbol{\theta}$.

The problem is assumed to be suitably regular in the sense that Laplace's method provides an approximation to the conditional Bayes factor given by

$$BF_{01}(\boldsymbol{Y}|\phi) \quad \approx \quad \frac{|\hat{\boldsymbol{I}}_n(\hat{\boldsymbol{\theta}}|\phi)|^{1/2}}{(2\pi)^{\nu/2}\pi(\hat{\boldsymbol{\theta}}|\phi)} e^{-\frac{1}{2}\|\boldsymbol{Z}(\phi)\|^2}, \tag{6}$$

as $n \to \infty$, where n is "sample size," ν is the dimension of $\boldsymbol{\theta}$,

$$\|\boldsymbol{Z}(\phi)\|^2 \quad = \quad 2l(\hat{\boldsymbol{\theta}}, \phi; \boldsymbol{Y}) - 2l(\boldsymbol{\theta}_0, \phi; \boldsymbol{Y}),$$

$\hat{\boldsymbol{\theta}}$ solves $\boldsymbol{\nabla} l(\hat{\boldsymbol{\theta}}, \phi; \boldsymbol{Y}) = \boldsymbol{0}$, and $\hat{\boldsymbol{I}}_n(\boldsymbol{\theta}|\phi) = -\boldsymbol{\nabla}^2 l(\boldsymbol{\theta}, \phi; \boldsymbol{Y})$, writing $\boldsymbol{\nabla}$ and $\boldsymbol{\nabla}^2$ to denote the gradient and Hessian operators with respect to $\boldsymbol{\theta}$. For example, the Laplace approximation (6) holds when $l(\boldsymbol{\theta}, \phi; \boldsymbol{Y}) + \log \pi(\boldsymbol{\theta}|\phi)$ is concave in $\boldsymbol{\theta}$, at least locally near its maximum value; see Tierney and Kadane (1986) for alternative conditions. An additional assumption is that, conditionally given ϕ, data generated under model $\mathcal{M}_1$ will induce

$$\hat{\boldsymbol{I}}_n(\hat{\boldsymbol{\theta}}|\phi) \approx \boldsymbol{I}_n(\boldsymbol{\theta}|\phi) \quad \text{and} \quad \pi(\hat{\boldsymbol{\theta}}|\phi) \to \pi(\boldsymbol{\theta}|\phi) > 0, \tag{7}$$

as $n \to \infty$, for a matrix-function $\boldsymbol{I}_n(\boldsymbol{\theta}|\phi)$, which is an asymptotic conditional Fisher information matrix. The same asymptotic properties are assumed for data generated under $\mathcal{M}_0$, except $\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

2.1 Specifying neutral data

Suppose it is possible to sensibly formulate a full-rank analogue $\boldsymbol{I}_0(\boldsymbol{\theta}|\phi)$ to $\boldsymbol{I}_n(\boldsymbol{\theta}|\phi)$ that is to represent the case where n is set to its minimum value. For example, this quantity may be the units in the rate of growth, $\boldsymbol{I}_0(\boldsymbol{\theta}|\phi) \approx n^{-1}\boldsymbol{I}_n(\boldsymbol{\theta}|\phi)$, or it might be devised by substituting into $\boldsymbol{I}_n(\boldsymbol{\theta}|\phi)$ the minimal sample-size information thought necessary to begin to understand the phenomenon under study. For insight into this idea, consider that when $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ is an independent and identically distributed sample, Fisher information is $\boldsymbol{I}_n(\boldsymbol{\theta}|\phi) = n\boldsymbol{I}_0(\boldsymbol{\theta}|\phi)$, which identifies the quantity $\boldsymbol{I}_0(\boldsymbol{\theta}|\phi)$ explicitly. The simple Gaussian case discussed in Section 1 has this set up, and yields $I_n(\theta) = n$ and $I_0(\theta) = 1$. Nevertheless, in multi-sample and other scenarios the notion of "sample size" is complex and the formulation of $\boldsymbol{I}_0(\boldsymbol{\theta}|\phi)$ requires special care. This is a familiar complication that also arises in applications of the Schwarz criterion, and is discussed further in Section 4.

Assuming it is possible to formulate a suitable $\boldsymbol{I}_0(\boldsymbol{\theta}|\phi)$, a *scaled* unit-information prior takes the form

$$\pi(\boldsymbol{\theta}|\phi) \quad = \quad (2\pi\tau^2)^{-\nu/2}|\boldsymbol{I}_0(\boldsymbol{\theta}_0, \phi)|^{1/2}h(\boldsymbol{\theta}|\phi), \tag{8}$$

where $\tau^2 > 0$ is the scale parameter, and

$$h(\boldsymbol{\theta}|\phi) \quad = \quad f\left(\frac{1}{2\tau^2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T\boldsymbol{I}_0(\boldsymbol{\theta}_0, \phi)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right), \tag{9}$$

for some function f. Within this family, the unit-information prior is defined at the setting $\tau^2 = 1$, at which the "amount of information in the prior on [the parameter] is equal to the amount of information about [the parameter] contained in one observation," by Kass and Wasserman's (1995, p. 929) characterization.

A prior of the form (8) is assumed throughout this article. Although this puts a definite restriction on the scope of proposed method's applicability, it is nevertheless the case that the scaled unit-information priors are suitable in a wide range of data-analysis scenarios. To use these priors, the analyst must typically attend carefully to identifying a parameter-transformation such that the prior remains meaningful even when τ^2 is large. The function f is often chosen to reflect a Gaussian or Cauchy prior, the latter touching

on the recommendation of Jeffreys (1961), or may be determined by integrating over the hyper-prior in a hierarchical formulation. Possible extension of present ideas to other priors is discussed in Section 4.

The proposed calibration rule in this scenario is to specify $\tilde{\boldsymbol{Y}}$ so that the neutral-data version of the conditional Bayes factor becomes

$$BF_{01}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi}) \quad = \quad \tau^{\nu}, \tag{10}$$

the same rule applied to the simple Gaussian case of Section 1, except here it is formulated conditionally, given $\boldsymbol{\phi}$, for a ν-dimensional parameter. As before, the setting $\tau^2 = 1$ reflects the desired property that neutral data are to yield neutral evidence under a unit-information prior. The precise form τ^{ν} reflects the typical scaling properties of a Bayes factors, for whenever a prior density is from a scale family, $\pi(\boldsymbol{\theta}) = \tau^{-\nu}\pi^*(\boldsymbol{\theta}/\tau)$, the Bayes factor has $BF_{01} \approx \tau^{\nu} BF_{01}^*$ as $\tau \to \infty$, where BF_{01}^* is defined at $\tau^2 = 1$.

2.2 Approximations and connections

Application of the Laplace approximation (6), together with (1) and (10), provides an approximation to the neutral-data comparison given by

$$NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \quad \approx \quad \frac{|\hat{\boldsymbol{I}}_n(\hat{\boldsymbol{\theta}},\boldsymbol{\phi})|^{1/2}}{|\boldsymbol{I}_0(\boldsymbol{\theta}_0,\boldsymbol{\phi})|^{1/2}} \frac{e^{-\frac{1}{2}\|\boldsymbol{Z}(\boldsymbol{\phi})\|^2}}{h(\hat{\boldsymbol{\theta}}|\boldsymbol{\phi})}. \tag{11}$$

Kass and Wasserman (1995) derive a similar approximation to the Bayes factor, under the setting $\tau^2 = 1$, and explore its asymptotic properties when $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + O(n^{-1/2})$. Applying the same assumption within (11) implies $h(\hat{\boldsymbol{\theta}}|\boldsymbol{\phi}) \approx f(\boldsymbol{0})$, and, by (7), the subsequent approximation $NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \approx \exp S_{01}(\boldsymbol{Y}|\boldsymbol{\phi})$ as $n \to \infty$, having defined the modified Schwarz criterion

$$S_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \quad = \quad -\frac{1}{2}\|\boldsymbol{Z}(\boldsymbol{\phi})\|^2 + \log \frac{|\hat{\boldsymbol{I}}_n(\hat{\boldsymbol{\theta}},\boldsymbol{\phi})|^{1/2}}{|\boldsymbol{I}_0(\boldsymbol{\theta}_0,\boldsymbol{\phi})|^{1/2}} - \log f(\boldsymbol{0}). \tag{12}$$

In the case where $\boldsymbol{\phi}$ is absent and $\boldsymbol{I}_n(\boldsymbol{\theta}) = n\boldsymbol{I}_0(\boldsymbol{\theta})$, the formula (12) exactly matches Kass and Wasserman's (1995) modified Schwarz criterion, in which $f(\boldsymbol{0})$ adjusts for a non-Gaussian prior.

In the development of neutral-data comparisons, asymptotic behavior as $\tau^2 \to \infty$ is a more central concern than asymptotic behavior as $n \to \infty$. By this perspective, it is interesting to observe that $h(\boldsymbol{\theta}|\boldsymbol{\phi}) \to f(\boldsymbol{0})$ as $\tau^2 \to \infty$, hence (11) shows that, when n is large and τ^2 is *very* large, the neutral-data comparison $NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi})$ is very nearly the exponentiated Schwarz criterion in (12).

Under present assumptions, the approximation $NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \approx \exp S_{01}(\boldsymbol{Y}|\boldsymbol{\phi})$ as $n \to \infty$ is accurate under $\mathcal{M}_0$, but not under $\mathcal{M}_1$, although, from a practical point of view, the inaccuracy is small relative to any moderate level of support provided by $NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi})$ for $\mathcal{M}_1$. This is illustrated in Section 3.1, below. The same pattern of inaccuracy arises when approximating the conditional Bayes factor at $\tau^2 = 1$ according to $BF_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \approx \exp S_{01}(\boldsymbol{Y}|\boldsymbol{\phi})$ as $n \to \infty$, which is the basis of criticisms made in Berger and Pericchi (1996) and Moreno *et al.*, (1999) of the Schwarz criterion. The *exact* neutral-data comparison (1) is a valid Bayesian procedure on its own, which need not be interpreted as an approximate Bayes factor, and so avoids a parallel criticism. The limiting neutral-data comparison, as $\tau^2 \to \infty$, is

$$NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \quad \to \quad \left\{(2\pi)^{-\nu/2}|\boldsymbol{I}_0(\boldsymbol{\theta}_0,\boldsymbol{\phi})|^{1/2} f(\boldsymbol{0}) \int e^{l(\boldsymbol{\theta},\boldsymbol{\phi};\boldsymbol{Y}) - l(\boldsymbol{\theta}_0,\boldsymbol{\phi};\boldsymbol{Y})} d\boldsymbol{\theta}\right\}^{-1}, \tag{13}$$

provided the integral exists. It would be consistent with the above arguments to regard the limit in (13) as the neutral-data comparison under a completely non-informative and possibly improper prior, or as a "small sample" version of the exponentiated Schwarz criterion. Either way, the limit in (13), too, avoids any criticism of inaccuracy as $n \to \infty$, hence resolves the shortcomings of $\exp S_{01}(\boldsymbol{Y}|\boldsymbol{\phi})$.

3 Demonstrations on example data

In this section, the proposed methodology derived from the rule (10) is demonstrated in two example data-analyses. Prior to that discussion, it is helpful to comment on computations, and on neutral-data comparisons' connection to the prior and posterior model probabilities.

In the example analyses, the unconditional version of a Bayes factor or neutral-data comparison is calculated by the following steps. First, the conditional version (5) is converted into conditional posterior model probabilities, and then to unconditional posterior model probabilities using the formula

$$P[\mathcal{M}_0|\boldsymbol{Y}] \quad = \quad \left\{1 + \frac{\int P[\mathcal{M}_1|\boldsymbol{Y},\boldsymbol{\phi}]\pi_0(\boldsymbol{\phi}|\boldsymbol{Y})d\boldsymbol{\phi}}{\int P[\mathcal{M}_0|\boldsymbol{Y},\boldsymbol{\phi}]\pi_1(\boldsymbol{\phi}|\boldsymbol{Y})d\boldsymbol{\phi}}\right\}^{-1}, \tag{14}$$

having written $\pi_0(\boldsymbol{\phi}|\boldsymbol{Y})$ and $\pi_1(\boldsymbol{\phi}|\boldsymbol{Y})$ for the model-specific posterior densities of the nuisance parameter. The final step is to convert the unconditional posterior model probabilities to the desired unconditional assessment.

In order to carry out these steps, it is necessary to specify prior model probabilities, which are not required for calculating a Bayes factor or neutral-data comparison, but they are required for calculating a posterior model probability. An important relationship is

$$\rho_{01}(\boldsymbol{\phi}) = \tilde{\rho}_{01}(\boldsymbol{\phi})/BF_{01}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi}), \tag{15}$$

where $\rho_{01}(\boldsymbol{\phi}) = P[\mathcal{M}_0|\boldsymbol{\phi}]/P[\mathcal{M}_1|\boldsymbol{\phi}]$ and $\tilde{\rho}_{01}(\boldsymbol{\phi}) = P[\mathcal{M}_0|\tilde{\boldsymbol{Y}},\boldsymbol{\phi}]/P[\mathcal{M}_1|\tilde{\boldsymbol{Y}},\boldsymbol{\phi}]$. In the example analyses, the reported Bayes factors are calculated under the setting $\rho_{01}(\boldsymbol{\phi}) = 1$; the reported neutral-data comparisons are calculated under the setting $\tilde{\rho}_{01}(\boldsymbol{\phi}) = 1$, which calibrates $\rho_{01}(\boldsymbol{\phi})$ through formula (15). Refer to Spitzner (2011) for an interpretation of these settings. As it turns out, the analyst need not actually solve $\rho_{01}(\boldsymbol{\phi})$ from $\tilde{\rho}_{01}(\boldsymbol{\phi})$, but can instead work with the relationships,

$$P[\mathcal{M}_0|\boldsymbol{Y},\boldsymbol{\phi}]/P[\mathcal{M}_1|\boldsymbol{Y},\boldsymbol{\phi}] = \rho_{01}(\boldsymbol{\phi})BF_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) = \tilde{\rho}_{01}(\boldsymbol{\phi})NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi}). \tag{16}$$

The rightmost formula in (16) is especially convenient when working with extremely vague priors, since, in that case, $\rho_{01}(\boldsymbol{\phi})$ is near zero when $\tilde{\rho}_{01}(\boldsymbol{\phi})$ is of moderate size, making the middle formula difficult to use.

Integration in (14) is carried out numerically by averaging over model-specific MCMC-generated samples. In every set of data-analysis results presented below, the number of iterations is at least one million, yielding a very high level of simulation accuracy.

3.1 The Behrens-Fisher problem

The first example demonstrates the proposed methodology in the context of the Behrens-Fisher problem, using the "yarn strength" data from Box and Tiao (1992, ex. 2.5.4). The Behrens-Fisher problem is a simple, classic setup that has been studied by many authors, frequentist and Bayesian; it is curious for the complications to classical methodology that its nuisance parameters introduce. The problem involves two data vectors, $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$, which represent measurements drawn from independent samples of respective size n_1 and n_2. Box and Tiao's data describe measurements of yarn breaking-strength from samples of size $n_1 = 20$ and $n_2 = 12$, with respective sample means $\bar{Y}_1 = 50$ and $\bar{Y}_2 = 55$, and sample variances $s_1^2 = 12$ and $s_2^2 = 40$. The model $\mathcal{M}_0$ puts $\boldsymbol{Y}_i|\mu,\sigma_i^2 \sim G(\mu\boldsymbol{1},\sigma_i^2\boldsymbol{I}_{n_i})$ and $\mathcal{M}_1$, puts $\boldsymbol{Y}_i|\mu_i,\sigma_i^2 \sim G(\mu_i\boldsymbol{1},\sigma_i^2\boldsymbol{I}_{n_i})$.

Although formula (8) defines a scaled unit-information prior from Fisher information, it is possible in this problem to identify such a prior from more direct arguments. Observe that, under $\mathcal{M}_1$,

$$\left.\frac{n_1\bar{Y}_1/\sigma_1^2 + n_2\bar{Y}_2/\sigma_2^2}{n_1/\sigma_1^2 + n_2/\sigma_2^2}\right| \sigma_1^2,\sigma_2^2 \quad \sim \quad G\left(\frac{n_1\mu_1/\sigma_1^2 + n_2\mu_2/\sigma_2^2}{n_1/\sigma_1^2 + n_2/\sigma_2^2}, \frac{1}{n_1/\sigma_1^2 + n_2/\sigma_2^2}\right) \tag{17}$$

and, independently,

$$\left.\frac{\bar{Y}_1 - \bar{Y}_2}{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right| \sigma_1^2,\sigma_2^2 \quad \sim \quad G\left(\frac{\mu_1 - \mu_2}{\sigma_1^2/n_1 + \sigma_2^2/n_2}, \frac{1}{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right). \tag{18}$$

In light of these formulas, Kass and Wasserman's (1995) characterization of unit-information is easily adapted to reflect equality of *variance* in the prior to the variance associated with one observation *from each sample*. Applying this idea to (17) and (18) motivates the transformation

$$\mu = \frac{\mu_1/\sigma_1^2 + \mu_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2} \quad \text{and} \quad \theta = \frac{\mu_1 - \mu_2}{\sigma_1^2 + \sigma_2^2}, \tag{19}$$

and identifies the corresponding conditional scaled unit-information priors

$$\mu|\sigma_1^2,\sigma_2^2 \sim G\left(0, \frac{\tau^2}{1/\sigma_1^2 + 1/\sigma_2^2}\right) \quad \text{and} \quad \theta|\sigma_1^2,\sigma_2^2 \sim G\left(0, \frac{\tau^2}{\sigma_1^2 + \sigma_2^2}\right),$$

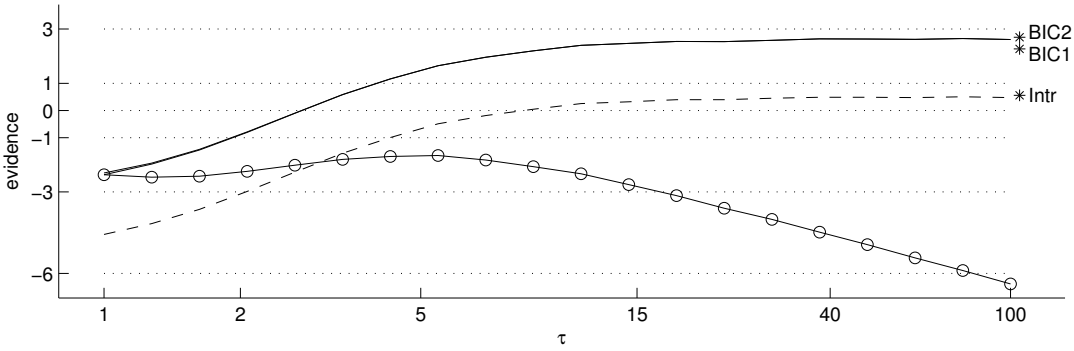

Figure 1: *Evidence assessments on Box and Tiao's yarn-strength data for τ between 1 and 100, plotted on a standard scale of evidence. The two (almost entirely overlapping) solid lines mark the default neutral-data comparisons and an approximation; the solid line with circles mark the Bayes factor; and the dashed line mark neutral-data comparisons calibrated to $\tau/3$.*

where τ is the scale parameter. The prior on the variance parameters is taken to specify independent scaled inverse-chi-square distributions, $\lambda/\sigma_1^2 \sim \chi^2_\kappa$ and $\lambda/\sigma_2^2 \sim \chi^2_\kappa$.

Under the transformation (19), the model $\mathcal{M}_1$ is re-parameterized to $\boldsymbol{Y}_1|\theta, \boldsymbol{\phi} \sim G((\mu + \sigma_1^2\theta)\mathbf{1}, \sigma_1^2 \boldsymbol{I}_{n_1})$ and $\boldsymbol{Y}_2|\theta, \boldsymbol{\phi} \sim G((\mu - \sigma_2^2\theta)\mathbf{1}, \sigma_2^2 \boldsymbol{I}_{n_2})$, having set $\boldsymbol{\phi} = (\mu, \sigma_1^2, \sigma_2^2)$. The model $\mathcal{M}_0$ is identified by the setting $\theta = \theta_0 = 0$. The relevant conditional Bayes factor is

$$BF_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \quad = \quad \left(1 + \tau^2 \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)^{1/2} \exp\left\{-\frac{1}{2} w(\sigma_1^2, \sigma_2^2) Z(\boldsymbol{\phi})^2\right\}, \tag{20}$$

where

$$Z(\boldsymbol{\phi})^2 = \frac{\{n_1(\bar{Y}_1 - \mu) + n_2(\bar{Y}_2 - \mu)\}^2}{n_1\sigma_1^2 + n_2\sigma_2^2} \quad \text{and} \quad w(\sigma_1^2, \sigma_2^2) = \frac{\tau^2(n_1\sigma_1^2 + n_2\sigma_2^2)/(\sigma_1^2 + \sigma_2^2)}{1 + \tau^2(n_1\sigma_1^2 + n_2\sigma_2^2)/(\sigma_1^2 + \sigma_2^2)},$$

and the corresponding neutral-data comparison is

$$NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \quad = \quad \left(\frac{1}{\tau^2} + \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)^{1/2} \exp\left\{-\frac{1}{2} w(\sigma_1^2, \sigma_2^2) Z(\boldsymbol{\phi})^2\right\}. \tag{21}$$

Though not specifically needed, it is straightforward to solve (10) in order to deduce an explicit setting of neutral data, whose analogue to $Z(\boldsymbol{\phi})^2$ is

$$\tilde{Z}(\boldsymbol{\phi})^2 \quad = \quad \frac{1}{w(\sigma_1^2, \sigma_2^2)} \log\left(\frac{1}{\tau^2} + \frac{n_1\sigma_1^2 + n_2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right).$$

Figure 1 displays unconditional assessments calculated from the neutral-data comparison (21), in several configurations, together with results obtained by the Bayes factor and other established procedures, including several calculated by Moreno *et al.* (1999) on the same data. Twenty values of the scale parameter are examined, across the range $1 \leq \tau \leq 100$, which indexes the horizontal axis of Figure 1. The prior variance parameters are set to $\kappa = \lambda = 0$, a standard non-informative setting, in every evaluation. The precise quantities that are plotted in Figure 1 are manifestations of the formula $2\log(P[\mathcal{M}_1|\boldsymbol{Y}]/P[\mathcal{M}_0|\boldsymbol{Y}])$, by which larger magnitudes indicate stronger evidence for $\mathcal{M}_0$ (if negative) or $\mathcal{M}_1$ (if positive); the strength of evidence is categorized into "positive," "strong," and "very strong" above the thresholds 3, 6, and 10, according to the scale proposed in Kass and Raftery (1995).

The results plotted in Figure 1 include those calculated from the approximation (11), which in the present problem evaluates to

$$NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \quad \approx \quad \left(\frac{n_1\sigma_1^2 + n_2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\right)^{1/2} \exp\left\{-\frac{1}{2} Z(\boldsymbol{\phi})^2 + \frac{1}{2}\left(\frac{\sigma_1^2 + \sigma_2^2}{\tau^2}\right)\hat{\theta}^2\right\}, \tag{22}$$

where $\hat{\theta} = \{n_1(\bar{Y}_1 - \mu) - n_2(\bar{Y}_2 - \mu)\}/(n_1\sigma_1^2 + n_2\sigma_2^2)$. Results from both (21) and (22) are plotted as solid lines in Figure 1. The graphs associated with these two statistics almost entirely overlap, indicating that

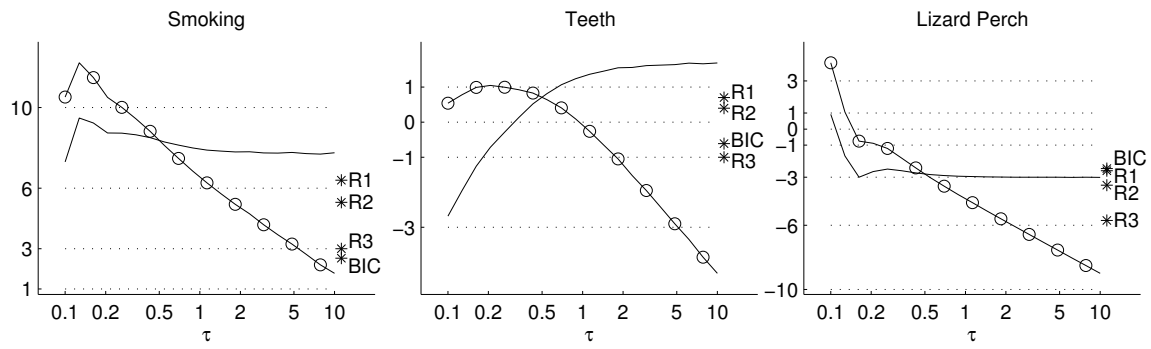

Figure 2: *Evidence assessments on Raftery's "Smoking," "Teeth," and "Lizard Perch" tables for τ between 0.1 and 10, plotted on a standard scale of evidence. The solid line marks the default neutral-data comparisons; the solid line with circles mark the Bayes factor. The asterisks labelled "R1," "R2," "R3," mark the assessments obtained under Raftery's prior at respective values 1, 1.65, and 5 of that prior's scale parameter.*

the inaccuracy of (22) to (21) is very small on these data; the graph associated with (22) is very slightly smaller. The solid line overlaid with circles in Figure 1 is calculated from the Bayes factor (20). As expected from its scaling properties, the evidence for $\mathcal{M}_1$ exhibited by the Bayes factor grows drastically weaker as τ increases beyond a certain value, while that exhibited by the neutral-data comparisons eventually stabilize.

Several assessments alluded to in previous discussion, but not addressed in detail, are also plotted in Figure 1. Results from two versions of the Schwarz criterion are marked by asterisks and labeled "BIC1" and "BIC2," each of which is calculated using a different *ad hoc* technique for handling the nuisance parameters and definition of "sample size." The value BIC1 is calculated in Moreno *et al.* (1999), and BIC2 is calculated using "MLE substitution," according to the formula

$$\hat{S}_{01}(\boldsymbol{Y}) \quad = \quad \frac{1}{2} Z(\hat{\boldsymbol{\phi}})^2 - \frac{1}{2} \log \left(\frac{n_1 \hat{\sigma}_1^2 + n_2 \hat{\sigma}_2^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2} \right),$$

where $\hat{\boldsymbol{\phi}} = (\hat{\mu}, \hat{\sigma}_1^2, \hat{\sigma}_2^2)$ is the maximum-likelihood value of $\boldsymbol{\phi}$ under $\mathcal{M}_1$. Refer to Bollen *et al.* (2012) for general discussion of such techniques. It is interesting that the results derived from neutral-data comparison tend to stabilize near those of the *ad hoc* Schwarz criteria; however, that pattern appears to be due, at least in part, to coincidence, as the pattern is quite different in the results of the second example analysis of Section 3.2, below.

Moreno *et al.* (1999) also calculate an intrinsic Bayes factor, whose result is marked in Figure 1 by an asterisk and labeled "Intr." As an experiment with using a family other than unit-information priors to define neutral data, a set of results is calculated from a modified calibration rule (10), by which τ^ν is revised to τ^ν/γ, where γ is selected so that the limiting neutral-data comparison approximately matches the intrinsic Bayes factor. The value identified here, by simple trial and error, is $\gamma = 3$; results corresponding to that setting are plotted as a dashed line in Figure 1. It is interesting that the theory of intrinsic priors suggests a different value, $\gamma = 2$, in the simple Gaussian case discussed in Section 1. The setting $\gamma = 3$ identified here likely reflects Moreno *et al.*'s (1999) specific handling of nuisance parameters in their construction of the intrinsic prior.

Such fiddling with the calibration rule (*i.e.*, introducing the constant γ) suggests intriguing possibilities for making an entirely *subjective* choice of neutral data. For instance, consider an exercise carried out in Moreno *et al.* (1999, sec. 3) in which the results of a number of Bayesian testing procedures are compared across an array of hypothetical data values; the authors conclude that the intrinsic Bayes factor shows "sensible discriminatory behavior" and is to be preferred. It is proposed that $\gamma = 1$, the value used throughout this article, be treated as a "default," and that an alternative, valid setting for γ might be determined by refining Moreno *et al.*'s exercise, based on the analyst's opinion of sensible behavior.

3.2 Log-linear models for the analysis of two-way tables

The second example demonstrates the proposed methodology in the analysis of two-way tables. The data are taken from Raftery (1993, sec 9.3) and consist of three 2×2 tables generated from separate experiments. Write $\boldsymbol{Y} = \{Y_{11}, Y_{12}, Y_{21}, Y_{22}\}$ to denote the data of an individual table, where Y_{jk} is the cell count of the j'th row and k'th column. The raw data are $\boldsymbol{Y} = \{32, 11, 60, 30\}$ for the "Smoking" experiment, $\boldsymbol{Y} = \{4, 16, 1, 21\}$ for the "Teeth" experiment, and $\boldsymbol{Y} = \{32, 11, 86, 35\}$ for the "Lizard Perch" experiment. See Raftery (1993) for sources and additional description.

The Y_{jk} are taken to be Poisson counts that are independent across the table cells. The models $\mathcal{M}_0$ and $\mathcal{M}_1$ are distinguished by the absence, in $\mathcal{M}_0$, or presence, in $\mathcal{M}_1$, of row-column interaction among the log-transformed Poisson means, $\eta_{jk} = \log E[Y_{jk}]$. The nuisance parameter, $\boldsymbol{\phi} = (\phi_1, \phi_2, \phi_3)$, collects the "free parameters," $\phi_1 = (\eta_{11}+\eta_{12}+\eta_{21}+\eta_{22})/2$, $\phi_2 = (\eta_{11}-\eta_{12}+\eta_{21}-\eta_{22})/2$, and $\phi_3 = (\eta_{11}+\eta_{12}-\eta_{21}-\eta_{22})/2$, which are orthonormal transformations of the η_{jk}. The parameter $\theta = (\eta_{11} - \eta_{12} - \eta_{21} + \eta_{22})/2$ identifies the magnitude of interaction, and is fixed at $\theta = \theta_0 = 0$ in model $\mathcal{M}_0$. The log-likelihood function is

$$l(\theta, \boldsymbol{\phi}; \boldsymbol{Y}) = \sum_{j,k} Y_{jk}\eta_{jk} - n(\theta, \boldsymbol{\phi}),$$

where $n(\theta, \boldsymbol{\phi}) = \sum_{j,k} e^{\eta_{jk}}$ gives the expected total count of table cells, in which the η_{jk} are understood as functions of θ and $\boldsymbol{\phi}$ by inverting the relationships identified above.

A suitable asymptotic framework treats $n(\theta_0, \boldsymbol{\phi})$ as "sample size," and considers asymptotic behavior as that quantity becomes arbitrarily large. It is furthermore assumed that each $E[Y_{jk}] = e^{\eta_{jk}}$ is asymptotically similar to $n(\theta_0, \boldsymbol{\phi})$, *i.e.*, each ratio is bounded above and bounded below above zero. This represents a "fixed marginal" scenario in which new measurements arrive independently to the table and fall into cells in proportions determined by the experimental phenomenon. The fixed marginal scenario is mechanically distinct from the "random marginal" scenario defined by Poisson counts, but it is easy to check that the respective likelihood functions are proportional, and so the scenarios are equivalent for purposes of inference. The dependence of sample size, $n = n(\theta_0, \boldsymbol{\phi})$, on a nuisance parameter is unconventional, but it nevertheless yields a Laplace approximation to the conditional Bayes factor (6), and is otherwise consistent with the framework of Section 2. The assumption of asymptotic similarity is necessary to be sure that the conditional maximum-likelihood value $\hat{\theta} \to \theta_0 = 0$, as $n(\theta_0, \boldsymbol{\phi}) \to \infty$, for data generated under $\mathcal{M}_0$.

It is straightforward to deduce that $\hat{I}_n(\hat{\theta}|\boldsymbol{\phi}) \approx I_n(\theta_0|\boldsymbol{\phi}) = n(\theta_0, \boldsymbol{\phi})/4$. The rate at which this quantity grows, relative to sample size, is $I_0(\theta_0|\boldsymbol{\phi}) = 1/4$, which is taken to define unit-information. The scaled unit-information prior applied here specifies $\theta \sim G(0, 4\tau^2)$, independently of $\boldsymbol{\phi}$. Similarly, the prior on $\boldsymbol{\phi}$ has independent $\phi_i \sim G(0, 4\tau^2)$. This is equivalent to specifying independent $\eta_{jk} \sim G(0, 4\tau^2)$ under model $\mathcal{M}_1$, and a constrained version of the same prior under model $\mathcal{M}_0$.

Analysis results on Raftery's count data, calculated at several settings of τ^2, are plotted in Figure 2. As in the example analysis of Section 3.1, the scale parameter is examined across of range of twenty values, $0.1 \leq \tau \leq 10$, which form the horizontal axis of each panel; as before, the quantities plotted are $2\log(P[\mathcal{M}_1|\boldsymbol{Y}]/P[\mathcal{M}_0|\boldsymbol{Y}])$, calculated from either a Bayes factor or neutral-data comparison, which indicate the strength of evidence for the model $\mathcal{M}_1$. Computations again rely on MCMC simulation, together with formula (14). In each panel of Figure 2, one sees the same pattern observed in the previous illustration, in which the Bayes factor exhibits increasingly stronger evidence for $\mathcal{M}_0$ at larger values of τ, while the neutral-data comparison stabilizes.

For reference, results associated with the Bayes factors calculated in Raftery (1993) are marked in each panel of Figure 2 by asterisks and labeled "R1," "R2," and "R3," which correspond to three values of a scale parameter for the class of priors used in those analyses. It is unsurprising that these plotted values are typically smaller than the values produced by neutral-data comparisons at large τ, since they presumably respond to scale in much the same way as the Bayes factors calculated here. A value derived from an *ad hoc* version of the Schwarz criterion also appears in each panel, marked by an asterisk and labeled "BIC." The calculation is made by the formula $\hat{S}_{01}(\boldsymbol{Y}) = l(\hat{\theta}, \hat{\boldsymbol{\phi}}_n; \boldsymbol{Y}) - l(\theta_0, \hat{\boldsymbol{\phi}}_n; \boldsymbol{Y}) - \frac{1}{2}\log N$, where $N = \sum_{jk} Y_{jk}$ and $\hat{\theta}$ and $\hat{\boldsymbol{\phi}}_n$ are maximum-likelihood values. It is interesting that the relative pattern in BIC is inconsistent across these examples: on the "Lizard Perch" data, the result based on BIC falls near those of the limiting neutral-data comparisons for large τ^2; in the other examples, the strength of evidence indicated by BIC for $\mathcal{M}_1$, relative to neutral-data comparisons, is substantially weaker.

4 Conclusions

A simple, intuitively reasonable calibration rule has been presented and explored for carrying out data analysis based on neutral-data comparisons. An implication of this rule is that it places a neutral-data comparison on a spectrum falling between a Bayes factor, formulated in a default configuration, and the exponentiated Schwarz criterion, highlighting that a neutral-data comparison is robust to modifications of the prior, but it is still sensitive to subjective knowledge. Exploration of the proposed methodology on existing data illustrates that neutral-data comparisons lead to reasonable conclusions, and produce values that are within the range of comparable assessments that have been calculated by other authors.

The concept of neutral data is a recent contribution to statistical theory, and is still under development as a tool for applied analysis. The present article offers a concrete guideline for specifying neutral data that is suitable for use with a widely applicable class of priors, the scaled unit-information priors. Extensions are certainly possible. For example, consider if $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ is such that $\boldsymbol{Y}_i \sim G(\boldsymbol{0}, \boldsymbol{\Sigma})$ under $\mathcal{M}_0$ and $\boldsymbol{Y}_i|\boldsymbol{\theta} \sim G(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with $\boldsymbol{\theta} \sim G(\boldsymbol{0}, \tau^2\boldsymbol{\Delta})$ under $\mathcal{M}_1$, independently across i. The Bayes factor is $BF_{01}(\boldsymbol{Y}) = |\boldsymbol{I} + \tau^2 n\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1/2}|^{1/2} \exp\{-\frac{1}{2}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\}$, where $\boldsymbol{Z} = n^{-1/2}\sum_i \boldsymbol{\Sigma}^{-1/2}\boldsymbol{Y}_i$ and $\boldsymbol{W} = \{\boldsymbol{I} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}^{1/2}/(\tau^2 n)\}^{-1}$. The scaling properties of the Bayes factor in this example suggests the modified calibration rule $BF_{01}(\tilde{\boldsymbol{Y}}) = \tau^{\nu}|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Delta}\boldsymbol{\Sigma}^{-1/2}|$, to replace (10). Such ideas are explored in Spitzner (2014), using ideas from Spiegelhalter and Smith (1982). A future investigation will develop concepts for understanding neutral data without recourse to asymptotic analysis, either as $n \to \infty$ or as $\tau^2 \to \infty$.

The proposed methodology makes heavy use of conditioning in order to deal with nuisance parameters. The example analysis of log-linear models in Section 3.2 highlights a particularly useful aspect of this approach, which is that it expands the concept of sample size to allow formulations that depend on nuisance parameters, as does the formulation of $n(\theta_0, \phi)$. The possibilities for meeting the requirements of intuition are widened by the added flexibility of parameter-specific formulations

A Motivation for neutral-data comparisons

Suppose the analyst is in the process of formulating a prior, and his or her attention is focused on the prior probabilities assigned to $\mathcal{M}_0$ and $\mathcal{M}_1$. As a check, he or she imagines a set of neutral data, $\tilde{\boldsymbol{Y}}$, which, matching the definition stated in Section 1, the analyst has come to think exhibits evidence no more in support of $\mathcal{M}_0$ than $\mathcal{M}_1$. On substituting $\tilde{\boldsymbol{Y}}$ into the posterior formula, he or she might expect to observe $P[\mathcal{M}_0] = P[\mathcal{M}_0|\tilde{\boldsymbol{Y}}]$, supposing that neutral data would fail to sway knowledge toward either model. However, the scale properties of Bayes factors imply that if $P[\mathcal{M}_0] = 1/2$, say, and the prior dispersion on $\mathcal{M}_1$ is much larger than that on $\mathcal{M}_0$ (and the imagined $\tilde{\boldsymbol{Y}}$ is not strongly tied to dispersion), then $P[\mathcal{M}_0|\tilde{\boldsymbol{Y}}]$ will be very nearly one. The analyst will, therefore, unexpectedly observe $P[\mathcal{M}_0] \neq P[\mathcal{M}_0|\tilde{\boldsymbol{Y}}]$. The effect of this observation is to create ambiguity over the choice of a *baseline* in weighing evidence. The Bayes factor, when written as a ratio of posterior to prior odds, $BF_{01}(\boldsymbol{Y}) = \{P[\mathcal{M}_0|\boldsymbol{Y}]/(1 - P[\mathcal{M}_0|\boldsymbol{Y}])\}/\{P[\mathcal{M}_0]/(1 - P[\mathcal{M}_0])\}$, is seen to use $P[\mathcal{M}_0]$ as its baseline. A neutral-data comparison makes the other choice, $P[\mathcal{M}_0|\tilde{\boldsymbol{Y}}]$, and weighs evidence as a ratio of posterior odds on observed data to that on neutral data, $NDC_{01}(\boldsymbol{Y}) = \{P[\mathcal{M}_0|\boldsymbol{Y}]/(1 - P[\mathcal{M}_0|\boldsymbol{Y}])\}/\{P[\mathcal{M}_0|\tilde{\boldsymbol{Y}}]/(1 - P[\mathcal{M}_0|\tilde{\boldsymbol{Y}}])\}$. It is straightforward to check that formula (1) is identical to this ratio.

REFERENCES

Berger, J. O., and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91:109- 122.

Bollen, K., Ray, S., Zavisca, J., and Harden, J. J. (2012), A comparison of Bayes factor approximation methods including two new methods, *Sociological Methods and Research*. In press.

Box, G. E. P., and Tiao, G. C. (1992), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Good, I. J. (1950), *Probability and the Weighing of Evidence*, London: Griffin.

Jeffreys, H. (1961), *Theory of Probability, 3rd ed.*, Oxford: Oxford University Press.

Kass, R. E., and Raftery, A. E., (1995) Bayes factors, *Journal of the American Statistical Association,* 90:773-795.

Kass, R. E., and Wasserman, L., (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion, *Journal of the American Statistical Association,* 90:928-934.

Lu, P. (2012) Calibrated Bayes factors for model selection and model averaging, *Ph.D. Dissertation*, The Ohio State University, Department of Statistics.

Moreno, E., Bertolino, F.,and Racugno, W. (1999), Default Bayesian analysis of the Behrens-Fisher problem, *Journal of Statistical Planning and Inference*, 81:323-333

O'Hagan, A. (1995), Fractional Bayes factors for model comparisons, *Journal of the Royal Statistical Society B*, 57:99-138.

Raftery, A. E., (1993) Approximate Bayes factors and accounting for model uncertainty in generalized linear models, Technical Report 255, University of Washington, Department of Statistics.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461-464.

Spiegelhalter, D. J., and Smith, A. F. M. (1982), Bayes factors for linear and log-linear models with vague prior information, *Journal of the Royal Statistical Society B,* 44:377-387.

Spitzner, D. J. (2011) Neutral-data comparisons for Bayesian testing, *Bayesian Analysis*, 6:603-638.

Spitzner, D. J. (2014) Adjusting for multiplicities in variable selection using neutral-data comparisons. Under review.

Tierney, L., and Kadane, J. B., (1986), Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association,* 81:82-86.

Weakliem, D. L. (1999), A critique of the Bayesian information criterion for model selection, *Sociological Methods & Research*, 27:359-397.