Submitted to the Brazilian Journal of Probability and Statistics

Subjective Bayesian testing using calibrated prior probabilities

Dan J. Spitzner[a]

[a] *University of Virginia, P. O. Box 400135, Charlottesville, VA 22904-4135, USA*

Abstract. This article proposes a calibration scheme for Bayesian testing that coordinates analytically-derived statistical performance considerations with expert opinion. In other words, the scheme is effective and meaningful for incorporating objective elements into subjective Bayesian inference. It explores a novel role for default priors as anchors for calibration rather than substitutes for prior knowledge. Ideas ae developed for use with multiplicity adjustments in multiple-model contexts, and to address the issue of prior sensitivity of Bayes factors. Along the way, the performance properties of an existing multiplicity adjustment related to the Poisson distribution are clarified theoretically. Connections of the overall calibration scheme to the Schwarz criterion are also explored. The proposed framework is examined and illustrated on a number of existing data sets related to problems in clinical trials, forensic pattern matching, and log-linear models methodology.

1 Introduction

This article examines the uses and impact of calibrating prior model probabilities in Bayesian testing (a.k.a. model choice) problems. The context emphasizes the value of expert opinion in statistical analysis, while at the same time stresses the importance of analytically-derived statistical properties. In conventional terms, what is proposed is a framework for fully coherent, *subjective* Bayesian testing that is equipped with interpretational guidance for incorporating *objective* elements into inference.

The examination makes central use of "default priors," a concept that Bayarri *et al.* (2012), succinctly characterizes as priors that "are not subjective priors, and are chosen conventionally based on the models being considered." They are traditionally aligned with the goals of objective Bayes inference, but this article breaks from that traditional context and motivation by proposing a use for default priors in the *calibration* of modeling and evidence assessment, rather than as a convenient substitute for prior information.

Keywords and phrases. hypothesis testing, model choice, Bayes factors, default priors, multiplicity, variable selection, Schwarz criterion

Inspiration for this inquiry is from two primary directions, multiplicity adjustments and prior sensitivity of Bayes factors, although the framework's potential impact is much broader. These two methodological contexts are currently and actively discussed in the literature, and while they are explored in this article primarily for demonstrating how the proposed calibration framework may be applied and interpreted, the resulting discussion also offers novel insights that advance each context's development.

Multiplicity adjustments are treated within the special case of variable selection: Suppose there are p independent variables, each to be categorized as a "selected" or "omitted." It is shown in Section 3, below, that a setting on ratios of prior model odds that induces asymptotic consistency under very weak assumptions is

$$P[M_s]/P[M_t] \propto \ \#\text{ omitted variables in } M_s \tag{1}$$

whenever the model M_s has exactly one more omitted variable than M_t. Such a prior is most closely related to the truncated Poisson priors studied in Womack *et al.* (2015), which is shown there to yield asymptotic consistency. It is a substantial modification of the beta-binomial prior, a more conventional discrete prior used in variable selection. For the latter, see, *e.g.*, Scott and Berger (2010), Wilson *et al.* (2010), and Castillo *et al.* (2015). The present examination contributes slightly to this line of inquiry by re-characterizing the assumptions for asymptotic consistency in terms of "ultra-high" dimensionality and rates of diminishing signal strength.

Inquiry into variable selection connects to the article's main objectives as follows. The proposed framework offers intellectual machinery for interpreting (1) as a calibration of the equal-weight setting, for which the prior odds between any two models is one. It is in this way that our ideas depart from the usual mode of reporting the results of an analysis that involves a default prior: they reflect a preference for assessing evidence relative to expert opinion–prior odds of one, say–rather than to (1), the detached prescription of a default. In other words, our ideas aim to retain the voice of the expert in assessing evidence, even when analytical considerations suggest incorporating into a statistical procedure elements such as asymptotic consistency that push against that voice.

As for prior sensitivity of Bayes factors, this article offers a novel interpretation of a technique for avoiding oversensitivity by jointly specifying the discrete and continuous portions of the prior. The technique is introduced in Robert (1993), and further developed in Spitzner (2011) and Dellaportas *et*

al. (2012). It is notable for offering a resolution to a certain "paradox" of inference that is commonly attributed to Lindley (1957), Bartlett (1957), and Jeffreys (1961). In this article, this joint-specification technique is presented as a calibration of prior probabilities.

The prior sensitivity issue is also valuable for motivating key insights of the proposed framework. Consider a very simple version of the Gaussian means problem, in which the target of inference is the mean, θ, of a random sample, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Assume a Gaussian model for data-generation, $Y_i|\theta \sim G(\theta, 1)$, and suppose the the "null" model M_0 constrains the mean to $\theta = 0$, while the "alternative" model M_1 has $\theta \sim G(0, \tau^2)$. The Bayes factor for M_0 *vs* M_1 is

$$BF_{01}(\boldsymbol{Y}) \quad = \quad (1 + \tau^2 n)^{1/2} \exp\{-\tfrac{1}{2} w_n Z^2\}, \tag{2}$$

where $Z = n^{1/2}\bar{Y}$, where $\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$, and $w_n = \tau^2 n/(1 + \tau^2 n)$. This follows from the general formula $BF_{01}(\boldsymbol{Y}) = \pi_0(\boldsymbol{Y})/\pi_1(\boldsymbol{Y})$, having written $\pi_s(\boldsymbol{Y})$ for the marginal density of $\boldsymbol{Y}$ under model M_s. Recall that the Bayes factor is conventionally interpreted to quantify "weight of evidence," by which larger values of $BF_{01}(\boldsymbol{Y})$ indicate stronger evidence for M_0, and smaller values indicate stronger evidence for M_1. (See, Kass and Raftery, 1995, for additional discussion of Bayes factors.)

A troublesome property of (2) is its unboundedness across the range of possible values of the prior scale parameter, τ, for $BF_{01}(\boldsymbol{Y})$ increases without bound as τ grows large. The implication is that as the expert may consider larger values of τ, beyond some hazy threshold the data's influence on weight of evidence becomes drastically overshadowed by the that of the prior. Upon being confronted with this property, Jeffreys (1961, p. 251), expressed reassurance by writing that "...the mere fact that it has been suggested that [the parameter] is zero corresponds to some presumption that it is fairly small." In other words, he suggests that unboundedness does not matter because the expert would naturally restrict themselves to a range of values for τ within which the Bayes factor behaves sensibly.

This article takes a different viewpoint, which is not of reassurance, but of concern that unboundedness discourages the participation of experts into inquiry: one imagines an unfortunate scene in which the expert, perhaps recruited into a study on promises of coherency in Bayesian inquiry, and accustomed to Bayesian estimation wherein statistical summaries converge to meaningful limit points as $\tau \to \infty$, becomes bewildered and discouraged at the inflexibility of what supposedly counts as meaningful prior knowledge

in Bayesian testing. The framework developed in this article is an effort to avoid this scene; it contributes to a broad goal for the development of data-analysis methodology of encouraging our experts to engage in inquiry and contribute their voice with full freedom of expression.

The article's ambitions are achieved in the following way. Default priors are removed to a peripheral role in testing, such that their only purpose is to identify an "anchor point" in the data space. The expert's prior is then calibrated against that point, combined with the observed data in the usual way, and converted into a calibrated Bayes factor. Some initial efforts in this direction are made in Spitzner (2011), whose ideas are built upon here and augmented with practical ideas for implementation. Theoretical evaluation and numerical demonstration of the resulting procedures explore its effectiveness at clarifying and interpreting evidence in multiple testing contexts, the generality of the approach in Gaussian and regular non-Gaussian models, the ease at which it allows formulations that depend on nuisance parameters, and its connections and non-connections to Schwarz's (1978) model-choice criterion.

The article's main ideas are developed in Section 2 in the context of null-*vs*-alternative model-comparison, M_0 *vs* M_1, in which each model is formulated from Gaussian distributions. Section 3 explores the formulation for multiple testing. Section 4 extends the formulation to regular non-Gaussian contexts. Demonstrations on existing data are presented in Section 5, and conclusions are made in Section 6. Proofs of the article's major mathematical assertions are placed in the appendix.

2 Main elements of the proposed framework

The main concepts and techniques of the proposed calibration framework are initially described in the context of a comparison between null and alternative models, M_0 *vs* M_1. In this setup, there is a parameter under test, $\boldsymbol{\theta}$, or "target" parameter, which is among the parameters of M_1 but not among those of M_0, where it has been constrained to a null value. Possibly, there is also a "nuisance" parameter $\boldsymbol{\phi}$, which is common to both models. Any nuisance parameter is treated as a quantity to be *conditioned* upon during analysis formulation, and integrated across when calculating analysis results. Accordingly, a prominent concept in the calibration framework is a conditional version of the Bayes factor,

$$BF_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \;=\; \frac{P[M_0|\boldsymbol{Y},\boldsymbol{\phi}]/P[M_1|\boldsymbol{Y},\boldsymbol{\phi}]}{\rho_{01}(\boldsymbol{\phi})} = \frac{\pi_0(\boldsymbol{Y}|\boldsymbol{\phi})}{\pi_1(\boldsymbol{Y}|\boldsymbol{\phi})}, \tag{3}$$

in which $\rho_{01}(\boldsymbol{\phi}) = P[M_0|\boldsymbol{\phi}]/P[M_1|\boldsymbol{\phi}]$ is conditional prior odds, and $\pi_s(\boldsymbol{Y}|\boldsymbol{\phi})$ is a marginal density for the data under model M_s, conditional on $\boldsymbol{\phi}$. Techniques for integrating (3) across models are illustrated in the application examples of Section 5.

2.1 Calibration to a default anchor in the data space

The proposed calibration framework is most readily formulated under the assumption that $\pi_1(\boldsymbol{\theta}|\boldsymbol{\phi})$, the conditional prior for $\boldsymbol{\theta}$ given $\boldsymbol{\phi}$, is embedded within a parametric family indexed by the prior parameter τ, whose value is specified from expert knowledge. Though many of the concepts proposed in this article are meaningful generally, the present framework is developed only for the case in which τ is a scale parameter. Such narrowing of focus is motivated by insights into this case put forward in Robert (1993), which call for the prior odds of M_0 to M_1 to reflect a contrast between the models in terms of high-probability regions of the target parameter: whereas the constraint placed on $\boldsymbol{\theta}$ under M_0 fixes those regions, any high-probability region associated with $\pi_1(\boldsymbol{\theta}|\boldsymbol{\phi})$ covers a wider range of values of $\boldsymbol{\theta}$ as the scale parameter increases. When this type of contrast is absent, calibration may be unnecessary, such as in conventional one-sided testing, or in scenarios where $\boldsymbol{\theta}$ is on a weak measurement scale. See also the discussion of the "device of imaginary results" in Section 2.3, below, which offers a concept that is generally applicable to determine whether calibration is called for, and how it may be achieved.

The initial conceptual step toward calibration is for the analyst to choose a default prior concept and use it to identify a particular value, $\tau = \tilde{\tau}$, to serve as the parameter's default setting. There is typically a variety of options to choose from, and we presume the analyst will choose their favorite. For example, in the Gaussian means problem that gives rise to (2), the "unit-information" default prior concept (see, Kass and Wasserman, 1995) would lead that analyst to select the value $\tilde{\tau} = 1$; the "intrinsic" default prior concept (see, Berger and Pericchi, 1996) would lead that analyst to select the value $\tilde{\tau} = \sqrt{2}$. Should the Gaussian prior, $\theta \sim G(0, \tau^2)$, be replaced with a Cauchy prior, $\theta \sim \text{Cauchy}(0, \tau)$, then Jeffreys (1961) recommended default setting of $\tilde{\tau} = 1$ would apply. It is not among this article's objectives to argue which default setting is "best." Instead, we assume the analyst has developed their own intuition as to a default prior concept that makes the most sense to them. As a starting point to the literature on default prior concepts, which is quite large, see *e.g.*, Jeffreys (1961), Zellner (1986), O'Hagan(1995), Kass and Wasserman (1995), Berger and Pericchi (1996), Ibrahim and Chen

(2000), Pérez and Berger (2002), Berger and Pericchi (2004), Liang *et al.* (2008), Casella *et al.* (2009), Bayarri *et al.* (2012), Moreno and Pericchi (2014), Fouskakis *et al.* (2017), and references therein.

Once a default value $\tilde{\tau}$ is identified, the next step is to use that value to locate an "anchor point" in the data space, against which prior probabilities and evidence are to be calibrated. This is defined as a point, $\boldsymbol{Y} = \tilde{\boldsymbol{Y}}$, such that

$$BF_{01}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi}) = 1 \quad \text{at} \quad \tau = \tilde{\tau}. \tag{4}$$

Note that $\boldsymbol{\phi}$ is omitted in the notation for $\tilde{\tau}$ and $\tilde{\boldsymbol{Y}}$, despite that both quantities are defined conditionally on that parameter. Non-uniqueness of $\tilde{\boldsymbol{Y}}$ is expected, but an easy remedy, described below in Sections 2.2 and 4.1, is available to accommodate this issue. Once a suitable $\tilde{\boldsymbol{Y}}$ is found, the default value, $\tilde{\tau}$, may be discarded.

The anchor point, $\tilde{\boldsymbol{Y}}$, is then used to formulate a calibrated Bayes factor. This is defined as

$$NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \;\; = \;\; \frac{P[M_0|\boldsymbol{Y},\boldsymbol{\phi}]/P[M_1|\boldsymbol{Y},\boldsymbol{\phi}]}{\tilde{\rho}_{01}(\boldsymbol{\phi})} = \frac{BF_{01}(\boldsymbol{Y}|\boldsymbol{\phi})}{BF_{01}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi})}, \tag{5}$$

where $\tilde{\rho}_{01}(\boldsymbol{\phi}) = P[M_0|\tilde{\boldsymbol{Y}},\boldsymbol{\phi}]/P[M_1|\tilde{\boldsymbol{Y}},\boldsymbol{\phi}]$, and all quantities other than $\tilde{\boldsymbol{Y}}$ are calculated at the expert's setting for τ. The label at the left in (5) is shorthand for "neutral-data comparison," a concept developed in Spitzner (2011) that interprets $\tilde{\boldsymbol{Y}}$ as "neutral" (imaginary) data. This interpretation is also suggested in criterion (4), for when the Bayes factor is one no evidence is exhibited more in support of one model than the other.

Observe from the middle expressions of (5) and (3) that a calibrated Bayes factor is a revision of the usual Bayes factor from a comparison of posterior to prior odds to a comparison of posterior odds on observed to neutral data. The rightmost expression in (5) is explicit in stating how the Bayes factor is calibrated relative to the anchor point. Section 2.3, below, offers a brief summary of the "neutral data" concepts developed in Spitzner (2011) that justify substituting (5) for (3) in evidence assessment.

The calibrated Bayes factor gives rise to a calibration of prior model probabilities in the following way. Combine the middle expressions of (5) and (3) to observe the relationships

$$P[M_0|\boldsymbol{Y},\boldsymbol{\phi}]/P[M_1|\boldsymbol{Y},\boldsymbol{\phi}] = \rho_{01}(\boldsymbol{\phi})BF_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) = \tilde{\rho}_{01}(\boldsymbol{\phi})NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi}). \tag{6}$$

Subsequently, the rightmost expressions in (5) and (6) combine to imply

$$\rho_{01}(\boldsymbol{\phi}) \quad = \quad \tilde{\rho}_{01}(\boldsymbol{\phi})/BF_{01}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi}), \tag{7}$$

which articulates the desired calibration.

The following is a summary of the steps of the approach just described. The assumption that $\pi_1(\boldsymbol{\theta}|\boldsymbol{\phi})$ is embedded in a scale family alludes to a set of *pre-calibration* steps: Step A, identify target and nuisance parameters; Step B, identify the scale parameter; and, Step C, if not obvious, derive a suitable mathematical framework for working with the model conditionally. These steps are aspects of model elicitation, and might be carried out in any analysis to gain insight into the inferential or computational framework. Subsequent to these steps in the proposed scheme are the following *calibration* steps: Step D, choose a default prior concept and set τ to a default value, $\tilde{\tau}$; Step E, find an "anchor point" $\tilde{\boldsymbol{Y}}$ by solving equation (4); Step F, discard $\tilde{\tau}$ and use $\tilde{\boldsymbol{Y}}$ to calculate the quantity $BF_{01}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi})$, the divisor in formula (5) for calibrating the Bayes factor, and in formula (7) for calibrating prior odds. These steps are referred to in later examples and discussion.

2.2 Example: The Gaussian means problem

The following examination of a multivariate version of the Gaussian means problem illustrates the concepts laid out above, and offers an approach to handling the non-uniqueness problem when the anchor point, $\tilde{\boldsymbol{Y}}$, is determined by the criterion (4).

Suppose a sample of n independent ν-dimensional measurements, $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$, is observed. The data are generated from Gaussian distributions, $\boldsymbol{Y}_i|\boldsymbol{\Sigma} \sim G(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, such that the mean parameter is restricted to $\boldsymbol{\theta} = \mathbf{0}$ under model M_0, but left unrestricted under M_1. The covariance matrix is treated as a nuisance parameter, $\boldsymbol{\phi} = \boldsymbol{\Sigma}$. Suppose further that the prior distribution under M_1 is such that $\boldsymbol{\theta}|\boldsymbol{\Sigma} \sim G(\mathbf{0}, \tau^2\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Delta}\boldsymbol{\Sigma}^{1/2})$, where $\boldsymbol{\Delta}$ is a symmetric, positive-definite matrix, which has been standardized so that its largest eigenvalue is one. It follows that the conditional Bayes factor for the model-comparison M_0 *vs.* M_1 is

$$BF_{01}(\boldsymbol{Y}|\boldsymbol{\Sigma}) \quad = \quad (\tau^2 n)^{\nu/2}|\boldsymbol{\Delta}|^{1/2}|\boldsymbol{W}|^{-1/2}\exp\{-\tfrac{1}{2}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\}, \tag{8}$$

having written $\boldsymbol{Z} = n^{1/2}\boldsymbol{\Sigma}^{-1/2}\bar{\boldsymbol{Y}}$, where $\bar{\boldsymbol{Y}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{Y}_i$, and $\boldsymbol{W} = \{\boldsymbol{I} + \boldsymbol{\Delta}^{-1}/(\tau^2 n)\}^{-1}$. This completes the pre-calibration Steps A, B, and C of Section 2.1.

Now suppose a default value $\tilde{\tau}$ is identified from the analyst's favorite default prior concept, and the equation of criterion (4) is solved. The solutions are succinctly written in terms of the neutral-data analogue $\tilde{\boldsymbol{Z}}$ to $\boldsymbol{Z}$ in (8), which substitutes $\tilde{\boldsymbol{Y}}$ for $\boldsymbol{Y}$ in defining that quantity. Each solution has $\tilde{\boldsymbol{Z}} = \sqrt{\tilde{c}}\tilde{\boldsymbol{W}}^{-1/2}\boldsymbol{u}$, where $\tilde{\boldsymbol{W}} = \{\boldsymbol{I} + \boldsymbol{\Delta}^{-1}/(\tilde{\tau}^2 n)\}^{-1}$, $\tilde{c} = \log\{(\tilde{\tau}^2 n)^{\nu}|\boldsymbol{\Delta}||\tilde{\boldsymbol{W}}|^{-1}\}$, and $\boldsymbol{u}$ is any unit-length ν-dimensional vector, $\|\boldsymbol{u}\| = 1$. It follows that

$$BF_{01}(\tilde{\boldsymbol{Y}}|\boldsymbol{\Sigma}) = \left\{(\tau^2 n)^{\nu}|\boldsymbol{\Delta}||\boldsymbol{W}|^{-1}\right\}^{1/2} \Big/ \left\{(\tilde{\tau}^2 n)^{\nu}|\boldsymbol{\Delta}||\tilde{\boldsymbol{W}}|^{-1}\right\}^{\frac{1}{2}\boldsymbol{u}^T\tilde{\boldsymbol{W}}^{-1/2}\boldsymbol{W}\tilde{\boldsymbol{W}}^{-1/2}\boldsymbol{u}}. \tag{9}$$

This completes the calibration Steps D, E, and F of Section 2.1. Our attention now turns to the issue of non-uniqueness of $\tilde{\boldsymbol{Y}}$.

The calibrated Bayes factor (5) is

$$NDC_{01}(\boldsymbol{Y}|\boldsymbol{\Sigma}) = \left\{(\tilde{\tau}^2 n)^{\nu}|\boldsymbol{\Delta}||\tilde{\boldsymbol{W}}|^{-1}\right\}^{\frac{1}{2}\boldsymbol{u}^T\tilde{\boldsymbol{W}}^{-1/2}\boldsymbol{W}\tilde{\boldsymbol{W}}^{-1/2}\boldsymbol{u}} \exp\{-\tfrac{1}{2}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\}, \tag{10}$$

Upon noting that $\boldsymbol{W} \to \boldsymbol{I}$ as $\tau \to \infty$, it is clear from (10) that $NDC(\boldsymbol{Y}|\boldsymbol{\Sigma})$ converges to a meaningful value for evidence assessment even when τ is set to a very large value,

$$NDC_{01}(\boldsymbol{Y}|\boldsymbol{\Sigma}) \approx \left\{(\tilde{\tau}^2 n)^{\nu}|\boldsymbol{\Delta}||\tilde{\boldsymbol{W}}|^{-1}\right\}^{\frac{1}{2}\boldsymbol{u}^T\tilde{\boldsymbol{W}}^{-1}\boldsymbol{u}} \exp\left\{-\tfrac{1}{2}\|\boldsymbol{Z}\|^2\right\} \text{ as } \tau \to \infty, \tag{11}$$

That is, the calibrated Bayes factor is bounded and therefore avoids oversensitivity to prior scale, as desired.

Nevertheless, the statistic (10) is generally unsuitable for implementation, due to the non-uniqueness of $\tilde{\boldsymbol{Y}}$, which varies with respect to $\boldsymbol{u}$. (An exception is the case $\boldsymbol{\Delta} = \boldsymbol{I}$, for which $\boldsymbol{u}^T\tilde{\boldsymbol{W}}^{-1/2}\boldsymbol{W}\tilde{\boldsymbol{W}}^{-1/2}\boldsymbol{u} = \{1+1/(\tilde{\tau}^2 n)\}/\{1+1/(\tau^2 n)\}$ does not vary.) Non-uniqueness may be handled in a satisfying way by appealing to large-sample properties: observe that both $\boldsymbol{W} \to \boldsymbol{I}$ and $\tilde{\boldsymbol{W}} \to \boldsymbol{I}$ as $n \to \infty$, so that $\boldsymbol{u}^T\tilde{\boldsymbol{W}}^{-1/2}\boldsymbol{W}\tilde{\boldsymbol{W}}^{-1/2}\boldsymbol{u} \to 1$, hence $BF_{01}(\tilde{\boldsymbol{Y}}|\boldsymbol{\Sigma})$ has a unique limiting value,

$$BF^* \quad = \quad (\tau/\tilde{\tau})^{\nu}. \tag{12}$$

Substituting (12) in place of (9), the calibrated Bayes factor is (5) is

$$NDC_{01}(\boldsymbol{Y}|\boldsymbol{\Sigma}) = (\tau^{\sim 2} n)^{\nu/2}|\boldsymbol{\Delta}|^{1/2}|\boldsymbol{W}|^{-1/2}\exp\{-^{1}{}_{2}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\}, \tag{13}$$

which, as with (11), avoids oversensitivity to prior scale, and is the ultimate form that is proposed for evidence assessment in problems of this sort. Arguments in Section 4, which develops the calibration framework for regular non-Gaussian models, motivate the use of a limiting value, an analogue to (12), as a convenient general means of resolving the non-uniqueness issue.

Note that the proposal is for large-sample analysis to be applied narrowly, only for the purpose of calibration. Still, given that it is used at all, one might argue for more a extensive use by which (10) is replaced with its approximation

$$NDC_{01}(\boldsymbol{Y}|\boldsymbol{\Sigma}) \;\approx\; (\tilde{\tau}^2 n)^{\nu/2}|\boldsymbol{\Delta}|^{1/2}\exp\left\{-\tfrac{1}{2}\|\boldsymbol{Z}\|^2\right\} \text{ as } n\to\infty. \quad (14)$$

This is not as desirable as (13), due to its excessive suppression of the expert's prior; *e.g.*, note the complete absence of the expert's scale parameter τ. It might, however, be regarded as a variation of the Bayes factor implied by Schwarz (1978) model-choice criterion, a connection that will be examined more carefully in Section 4.2.

In addition to inspiring a resolution to the non-uniqueness issue, the Gaussian means problem is also helpful for exploring the potential *impact* of non-uniqueness. For this purpose, let us focus on the formula (11) for the limiting value of the calibrated Bayes factor when τ is set to a very large value, and consider its range of values as the vector $\boldsymbol{u}$ varies. That range is determined from the eigenvalues of the matrix $\tilde{\boldsymbol{W}}^{-1}$, which, according to elementary matrix theory, bound the quadratic expression, $\boldsymbol{u}^T\tilde{\boldsymbol{W}}^{-1}\boldsymbol{u}$, appearing in the exponent of the formula's initial factor. Since $\boldsymbol{\Delta}$, which determines $\tilde{\boldsymbol{W}}^{-1}$, is standardized so that its largest eigenvalue is one, the bounds are found to be

$$1 + 1/(\tilde{\tau}^2 n) \le \boldsymbol{u}^T\tilde{\boldsymbol{W}}^{-1}\boldsymbol{u} \le 1 + \delta_1^{-1}/(\tilde{\tau}^2 n),$$

where δ_1 denotes the smallest eigenvalue of $\boldsymbol{\Delta}$. It is clear from these bounds that when $\boldsymbol{\Delta}$ is strongly ill-conditioned, hence δ_1 is very small, the calibrated Bayes factor varies widely across the solutions to criterion (4). In other words, potential impact of non-uniqueness could be quite severe.

As indicated, judicious application of large-sample asymptotics resolves this issue. However, should the reader not find that resolution compelling, a sensible second option would be to report the calibrated Bayes factor's range of possible values. As an attempt at a third option, it could be tempting to try to identify a preference for some particular value of $\boldsymbol{u}$ over all

others, but it is difficult to imagine any concept that would make such a preference meaningful. One seemingly convenient choice would be to prefer $\boldsymbol{u} = \boldsymbol{Y}/\|\boldsymbol{Y}\|$, thus aligning the anchor point to the observed data; yet, this would amount to a double-use of data, and would take us away from the subjective Bayesian framework for which the proposed methodology is intended. Ultimately, either of these alternative options would be challenging to implement in problems substantially more complex than the present Gaussian testing scenario, and neither are pursued here.

2.3 Interpretations and implications

The calibration concepts thus far described have a number of practical and interpretational implications. To set up the discussion of these aspects, it is helpful to first lay out the key arguments in Spitzner (2011) that motivate using the calibrated Bayes factor for assessing evidence.

The neutral-data concepts developed in Spitzner (2011) stem from a customized application of Good's (1950) "device of imaginary results," which uses "imaginary data" to check a candidate prior: Suppose the expert and analyst are in the process of eliciting a prior, and their attention is focused on the prior probabilities assigned to M_0 and M_1. As a check of some particular choice, the analyst imagines data, $\tilde{\boldsymbol{Y}}$, that would be characterized as "neutral," perhaps through a criterion like (4). Good's conceptual device is to apply the posterior calculation to $\tilde{\boldsymbol{Y}}$, and reflect upon whether the result produced is sensible. Upon doing so the analyst might expect to observe $P[M_0|\boldsymbol{\phi}] = P[M_0|\tilde{\boldsymbol{Y}}, \boldsymbol{\phi}]$, due to the neutrality of $\tilde{\boldsymbol{Y}}$. Nevertheless, the scale properties of a Bayes factor such as (2) or (8) would prevent this from happening, for $BF_{01}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi}) \to \infty$ as $\tau \to \infty$ implies $P[M_0|\tilde{\boldsymbol{Y}}, \boldsymbol{\phi}] \to 1$, regardless of $P[M_0|\boldsymbol{\phi}]$. (To see this, use the first equation in formula 6.) In other words, if τ is large, unless $P[M_0|\boldsymbol{\phi}]$ falls in an extreme range, the check fails.

The calibrated Bayes factor (5) is a resolution to this issue. To argue this point, Spitzner (2011) highlights the form of the Bayes factor (3) as a comparison of posterior to prior odds, and interprets the inequality $P[M_0|\boldsymbol{\phi}] \neq P[M_0|\tilde{\boldsymbol{Y}}, \boldsymbol{\phi}]$ as creating ambiguity over the choice of a *baseline* in weighing evidence. Whereas the Bayes factor (3) chooses prior odds, $\rho_{01}(\boldsymbol{\phi})$, as its baseline, the calibrated Bayes factor (5) makes the other choice, the quantity $\tilde{\rho}_{01}(\boldsymbol{\phi})$. In this way, the calibrated Bayes factor (5) is valid for assessing evidence.

Other implications of this argument are as follows. First, the association of $\tilde{\rho}_{01}(\boldsymbol{\phi})$ with the calibrated Bayes factor parallels that of $\rho_{01}(\boldsymbol{\phi})$ with the

usual Bayes factor, a parallel that is captured in the expressions of formula (6). What this means for practice is that the expert's opinion of odds is just as meaningfully assigned to $\tilde{\rho}_{01}(\boldsymbol{\phi})$ as $\rho_{01}(\boldsymbol{\phi})$. If the assignment is to $\tilde{\rho}_{01}(\boldsymbol{\phi})$ then (7) yields a meaningful calibration of $\rho_{01}(\boldsymbol{\phi})$. Such thinking admits the synthesis of subjective and objective elements alluded to at the start of Section 1. Moreover, if $\tilde{\rho}_{01}(\boldsymbol{\phi}) = 1$, then evidence is reported as a ratio of posterior model probabilities, to which the calibrated Bayes factor reduces, whose interpretation avoids certain criticisms of Bayes factors that are explored in Lavine and Schervish (1999).

Second, the rightmost expression in (6) is useful for computation when working with a very diffuse prior, since, in that case, $\rho_{01}(\boldsymbol{\phi})$ is near zero and therefore difficult to manage, whereas $\tilde{\rho}_{01}(\boldsymbol{\phi})$ is of a convenient size.

Third, the calibration formula (7) offers a precise conceptual mechanism for realizing Robert's (1993) proposal to jointly specify the discrete and continuous portions of the prior. Observe that the formula (7) implies that assigning a "large" value to τ induces a "small" value for $\rho_{01}(\boldsymbol{\phi})$, relative to $\tilde{\rho}_{01}(\boldsymbol{\phi})$, assuming the type of scaling behavior observed in (2) and (8). This is consistent with Robert's (1993) suggestion to choose $P[M_0|\boldsymbol{\phi}]$ in such a way that its ratio with the shrinking prior probability of some fixed "reasonable range" of values for $\boldsymbol{\theta}$ (*i.e.*, a compact subset in $\boldsymbol{\theta}$-space) is asymptotically constant as τ grows large. The proposed use of default priors makes this prescription precise.

3 Calibration in multiple testing

In the multiple-testing context there are multiple models M_s across $s \in S$, where S is a finite or countable index-set. For example, in variable selection with p variables, as in Section 1, the set S indexes the 2^p sub-models identified with the possible ways of selecting and omitting variables from consideration. Given some $s \in S$, the model M_s is defined from a subset $A_s \subset \{1, \ldots, p\}$, according to which the variables indexed by $i \in A_s$ are "omitted" in M_s and those with $i \notin A_s$ are "selected." For another example, a null-*vs*-alternative model-comparison falls trivially into the multiple-testing framework by setting $S = \{0, 1\}$.

3.1 Extending "null" *vs* "alternative" comparisons to multiple testing

In variable selection, a model-comparison, M_s *vs* M_t, such that $|A_s - A_t| = 1$ is "elementary" in the sense that it is a test of a single variable, the one that is simultaneously omitted in M_s and selected in M_t. The calibration concepts developed in Section 2 readily extend to this case in an obvious way: previous formulas are updated to

$$NDC_{st}(\boldsymbol{Y}|\boldsymbol{\phi}) \;=\; \frac{P[M_s|\boldsymbol{Y},\boldsymbol{\phi}]/P[M_t|\boldsymbol{Y},\boldsymbol{\phi}]}{\tilde{\rho}_{st}(\boldsymbol{\phi})} = \frac{BF_{st}(\boldsymbol{Y}|\boldsymbol{\phi})}{BF_{st}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi})} \qquad (15)$$

for the calibrated Bayes factor, updating (5), to

$$P[M_s|\boldsymbol{Y},\boldsymbol{\phi}]/P[M_t|\boldsymbol{Y},\boldsymbol{\phi}] = \rho_{st}(\boldsymbol{\phi})BF_{st}(\boldsymbol{Y}|\boldsymbol{\phi}) = \tilde{\rho}_{st}(\boldsymbol{\phi})NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \quad (16)$$

for posterior odds, updating (6), and to

$$\rho_{st}(\boldsymbol{\phi}) \;=\; \tilde{\rho}_{st}(\boldsymbol{\phi})/BF_{st}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi}), \qquad (17)$$

for the calibrated prior probability, updating (7), having written $\rho_{st}(\boldsymbol{\phi}) = P[M_s|\boldsymbol{\phi}]/P[M_t|\boldsymbol{\phi}]$ and $\tilde{\rho}_{st}(\boldsymbol{\phi}) = P[M_s|\tilde{\boldsymbol{Y}},\boldsymbol{\phi}]/P[M_t|\tilde{\boldsymbol{Y}},\boldsymbol{\phi}]$. In these formulas, the anchor point $\tilde{\boldsymbol{Y}}$ and nuisance parameter, $\boldsymbol{\phi}$, may be specific to the model-comparison M_s *vs* M_t.

Evidence assessment of a general (*i.e.*, possibly non-elementary) model-comparison M_s *vs* M_t is pieced together from that of relevant elementary model-comparisons in the following way. Set $j_1 = |A_s \cap A_t^c|$ and $j_2 = |A_s^c \cap A_t|$, and find a path $s = u_0, \ldots, u_{j_1}, u_{j_1+1}, \ldots, u_{j_1+j_2} = t \in S$, so that, for $1 \le r \le j_1$, the former model in $M_{u_{r-1}}$ *vs* M_{u_r} has exactly one more variable omitted, and, for $j_1 + 1 \le r \le j_1 + j_2$, the former model in $M_{u_{r-1}}$ *vs* M_{u_r} has exactly one more variable selected. At least one such path always exists. Applying (16), the ratio of posterior model probabilities is

$$\begin{aligned}
\frac{P[M_s|\boldsymbol{Y},\boldsymbol{\phi}]}{P[M_t|\boldsymbol{Y},\boldsymbol{\phi}]} &= \left\{\prod_{i=1}^{j_1} \frac{P[M_{u_{r-1}}|\boldsymbol{Y},\boldsymbol{\phi}]}{P[M_{u_r}|\boldsymbol{Y},\boldsymbol{\phi}]}\right\}\left\{\prod_{r=j_1+1}^{j_1+j_2} \frac{P[M_{u_{r-1}}|\boldsymbol{Y},\boldsymbol{\phi}]}{P[M_{u_r}|\boldsymbol{Y},\boldsymbol{\phi}]}\right\} \\
&= \left\{\prod_{r=1}^{j_1} \tilde{\rho}_{u_{r-1}u_r}(\boldsymbol{\phi})NDC_{u_{r-1}u_r}(\boldsymbol{Y}|\boldsymbol{\phi})\right\} \\
&\qquad \times \left\{\prod_{r=j_1+1}^{j_1+j_2} \tilde{\rho}_{u_{r-1}u_r}(\boldsymbol{\phi})\frac{1}{NDC_{u_ru_{r-1}}(\boldsymbol{Y}|\boldsymbol{\phi})}\right\}.
\end{aligned}$$

From this, the calibrated Bayes factor, $NDC_{st}(\boldsymbol{Y}|\boldsymbol{\phi})$, is calculated from the middle expression in (15).

3.2 Multiplicity adjustment

In variable selection, and other multiple-model contexts, a common argument put forward is that evidence assessment of a model-comparison M_s *vs* M_t, say, for $s, t \in S$, should take into account the presence of all models under consideration, not just those directly involved in the comparison. In other words, evidence for M_s *vs* M_t would be assessed differently if $S = \{s, t\}$ than if s and t are just two index-values among the 2^p index-values in S associated with variable selection in p variables. In the latter case, it is said that a "multiplicity adjustment" is applied to account for the presence of models other than those indexed by s and t.

A widely discussed approach to multiplicity adjustment in variable selection is to model the number of omitted variables through a binomial process. See, *e.g.*, Berry and Hochberg (1999), Scott and Berger (2010), Wilson *et al.* (2010), and Castillo *et al.* (2015). Denote by k_s the number of variables that are omitted (*i.e.*, $k_s = |A_s|$), and by ν_s the number that are selected (*i.e.*, $\nu_s = p - k_s$). The beta-binomial prior is typically formulated hierarchically according to $k_s|\xi \sim \text{binomial}(p, \xi)$ and $\xi \sim \text{beta}(\alpha, \beta)$, according to which a ratio of prior model probabilities, where M_s has one more variable omitted than M_t, becomes $P[M_s]/P[M_t] = (\beta + k_s - 1)(\alpha + \nu_s)$. The effect of this setting is to weight the omission of a variable when both M_s and M_t have many variables omitted, and weight the selection of a variable when both M_s and M_t have many variables selected. The setting (1), on the other hand, reflects that the mathematics for asymptotic consistency prescribe that only the omission of variables are to be weighted. These approaches are compared in Section 5.1 in a demonstration on example data. Decision-theoretic approaches to multiple-model testing are also available; see, *e.g.*, Müller, Parmigiani, and Rice (2007).

3.3 Asymptotic consistency in Gaussian variable selection

The present exploration of multiple-testing concepts focuses on the following simple, broadly applicable version of variable selection: each "variable" is identified with one of p independent sets of sample measurements, $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_p$; the i'th set is $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in})$ for independent Y_{ij}, which is associated with a parameter, θ_i. Underlying the setup is a fixed collection of "null" parameter values, $\boldsymbol{\theta}^0 = (\theta_i^0 : i = 1, \ldots, p)$. Should, under model M_s, the i'th variable be selected, its associated parameter, θ_i, is left "free." The free parameters are collected into $\boldsymbol{\theta}_s = (\theta_i : i \notin A_s)$, which forms the target parameter of model M_s. Should the i'th variable be omitted, its associated

parameter is set to its null value, $\theta_i = \theta_i^0$. Denote by ν_s the number of free parameters and by k_s the number of parameters set to null values.

Suppose now that, under model M_s, the data are generated according to $Y_{ij}|\boldsymbol{\theta}_s \sim G(\theta_i, 1)$ for $i \notin A_s$ and $Y_{ij} \sim G(\theta_i^0, 1)$ for $i \in A_s$. The prior has $\theta_i \sim G(\theta_i^0, \tau^2)$ for $i \notin A_s$, where τ is a scale parameter to be specified by the expert. Note, in this case, the absence of any nuisance parameter.

Having adopted model-specific prior distributions as stated, the setting (1) for prior model odds is achieved through the proposed calibration scheme at the values $\tilde{\tau} = k_s$ and $\tilde{\rho}_{st} = 1$. Under these settings, use of the limiting value (12) within the calibration formula (17) gives rise to (1) in the form $P[M_s]/P[M_t] = k_s/\tau$.

The following result establishes the desirability of this setting.

Theorem 1. *Suppose the data are generated from M_s;* i.e., *M_s is the "true" model. Suppose the prior model probabilities are such that if $u, v \in S$ satisfy $|A_u - A_v| = 1$ then $P[M_u]/P[M_v] = k_u/\tau$. Treat the number of variables as a function of n;* i.e., *$p = p_n$, and suppose there is a lower bound ξ_n such that $\xi_n \leq |\theta_i|$ for all $i \notin A_s$. Suppose further there are constants $a > 0$ and $b > 0$ such that $a < 1 - b$ and the following two conditions hold:*

*(*i*) $\log p_n = O(n^a)$, and*
*(*ii*) there is a $c > 0$ such that $\xi_n \geq cn^{-b}$.*

These conditions imply asymptotic consistency in the sense that

$$\liminf P[M_s|\boldsymbol{Y}] > 0 \text{ as } n \to \infty.$$

The conditions in Theorem 1 articulate the notion of "faint signals" ($\xi_n \geq cn^{-b}$) in "ultra-high dimensional" space ($\log p_n = O(n^a)$), and are very weak for establishing asymptotic consistency. An early reference to such conditions is Fan and Lv (2008), where they are used to evaluate a screening procedure known as "sure independence screening." Theorem 1 positions the Bayesian solution, using (1), among the few statistical procedures that are able to achieve asymptotic consistency under these conditions. For further discussion of this property and related procedures, see *e.g.*, Fan and Lv (2010) and Narissety and He (2014).

4 Calibration in non-Gaussian contexts

In a non-Gaussian context, calibration is readily formulated and practical to implement for models that arise from an exponential or other ex-

pansive parametric family. Suppose the testing problem is defined from a collection of models indexed by $s \in S$, and consider a particular model-comparison M_s *vs* M_t, for $s, t \in S$, which is nested in the sense that a parameter, $\boldsymbol{\theta}$, is present in M_t but set to a null value, $\boldsymbol{\theta}^0$, in M_s. A nuisance parameter, $\boldsymbol{\phi}$, *i.e.*, a parameter that is common to both models, may also be present. Because M_s and M_t are nested, a common log-likelihood function, $l_n(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{Y})$ is relevant to both models, which give rise to distinct marginal data-densities according to $\pi_s(\boldsymbol{Y}|\boldsymbol{\phi}) = l_n(\boldsymbol{\theta}^0, \boldsymbol{\phi}; \boldsymbol{Y})$ for model M_s and $\pi_t(\boldsymbol{Y}|\boldsymbol{\phi}) = \int \exp\{l_n(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{Y})\}\pi_t(\boldsymbol{\theta}|\boldsymbol{\phi})d\boldsymbol{\theta}$ for model M_t, where $\pi_t(\boldsymbol{\theta}|\boldsymbol{\phi})$ is the associated conditional prior.

An additional requirement is the presence of a "sample size" parameter, n, which may not strictly represent "the number of objects sampled," but is part of the framework in order to facilitate asymptotic analysis. Indeed, the example analyses of Section 5, below, illustrate the capacity of the proposed calibration framework to operate meaningfully even when sample size is ambiguously defined.

4.1 Finding an anchor point

The model-comparison M_s *vs* M_t is assumed to be suitably regular in the sense that Laplace's method provides an approximation to the conditional Bayes factor, analogous to (3), given by

$$BF_{st}(\boldsymbol{Y}|\boldsymbol{\phi}) \quad = \quad \frac{\pi_s(\boldsymbol{Y}|\boldsymbol{\phi})}{\pi_t(\boldsymbol{Y}|\boldsymbol{\phi})} \approx \frac{|\hat{\boldsymbol{I}}_n(\hat{\boldsymbol{\theta}}|\boldsymbol{\phi})|^{1/2}}{(2\pi)^{\nu/2}\pi_t(\hat{\boldsymbol{\theta}}|\boldsymbol{\phi})} e^{-\frac{1}{2}\|\boldsymbol{Z}(\hat{\boldsymbol{\theta}}|\boldsymbol{\phi})\|^2}, \tag{18}$$

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

$$\|\boldsymbol{Z}(\hat{\boldsymbol{\theta}}|\boldsymbol{\phi})\|^2 \quad = \quad 2l_n(\hat{\boldsymbol{\theta}}, \boldsymbol{\phi}; \boldsymbol{Y}) - 2l_n(\boldsymbol{\theta}^0, \boldsymbol{\phi}; \boldsymbol{Y}),$$

$\hat{\boldsymbol{\theta}}$ solves $\boldsymbol{\nabla} l_n(\hat{\boldsymbol{\theta}}, \boldsymbol{\phi}; \boldsymbol{Y}) = \boldsymbol{0}$, and $\hat{\boldsymbol{I}}_n(\boldsymbol{\theta}|\boldsymbol{\phi}) = -\boldsymbol{\nabla}^2 l_n(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{Y})$, writing $\boldsymbol{\nabla}$ and $\boldsymbol{\nabla}^2$ to denote the gradient and Hessian operators with respect to $\boldsymbol{\theta}$. For instance, the Laplace approximation (18) holds when $l_n(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{Y}) + \log \pi(\boldsymbol{\theta}|\boldsymbol{\phi})$ is concave in $\boldsymbol{\theta}$, at least locally near its maximum value; see Tierney and Kadane (1986) for alternative conditions.

Within this framework, the following simple approximation result offers a straightforward implementation of the proposed calibration scheme.

Theorem 2. *Suppose the approximation (18) holds, the conditional prior on* $\boldsymbol{\theta}$ *given* $\boldsymbol{\phi}$ *is from a scale family,* $\pi_t(\boldsymbol{\theta}|\boldsymbol{\phi}) = \tau^{-\nu}\pi^*(\boldsymbol{\theta}/\tau|\boldsymbol{\phi})$*, where* $\pi^*(\boldsymbol{\theta}/\tau|\boldsymbol{\phi})$

is finite and nonzero at $\boldsymbol{\theta} = \boldsymbol{\theta}^0$*, and* $|\hat{\boldsymbol{I}}_n(\boldsymbol{\theta}|\boldsymbol{\phi})| \to \infty$ *as* $n \to \infty$*, for any* $\boldsymbol{\theta}$*. Suppose further there is an anchor point* $\tilde{\boldsymbol{Y}}$ *that satisfies*

$$BF_{st}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi}) = 1 \quad at \quad \tau = \tilde{\tau}, \tag{19}$$

where $\tilde{\tau}$ *is a "default" value of the scale parameter. It follows that the Bayes factor calculated at the anchor point has*

$$BF_{st}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi}) \approx (\tau/\tilde{\tau})^{\nu} \quad as \quad n \to \infty. \tag{20}$$

The property deduced in Theorem 2 parallels and extends the behavior observed in the Gaussian means problem of Section 2.2, by confirming that the convergence of the Bayes factor to (12), when evaluated at the anchor point, is a general property. As we have seen, this type of result is important for resolving the potential conceptual complication that arises when a solution to (19) is not unique, for it indicates that any two distinct anchor-point solutions would yield nearly the same calibration. Even when the solution to (19) is unique, the target value (12) may still be important practically for simplifying how calibration would be implemented, for if (19) is hard to solve, and the analyst is willing to accept an approximate calibration, then they may simply work with the calibrated Bayes factor given by

$$NDC_{st}(\boldsymbol{Y}|\boldsymbol{\phi}) \quad = \quad BF_{st}(\boldsymbol{Y}|\boldsymbol{\phi})/BF^* = (\tilde{\tau}/\tau)^{\nu} BF_{st}(\boldsymbol{Y}|\boldsymbol{\phi}), \tag{21}$$

where BF^* is the limit point to $BF_{st}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi})$ identified in (12) and (20).

4.2 Connections to the Schwarz criterion

Kass and Wasserman (1995) put forward a description of the Schwarz criterion as an approximation to the Bayes factor that is formulated from a unit-information prior. An interesting connection to the calibrated Bayes factor formula (21) is drawn as follows.

Suppose that $\hat{\boldsymbol{I}}_n(\hat{\boldsymbol{\theta}}|\boldsymbol{\phi}) \approx \boldsymbol{I}_n(\boldsymbol{\theta}|\boldsymbol{\phi})$ as $n \to \infty$, for an asymptotic conditional Fisher information matrix $\boldsymbol{I}_n(\boldsymbol{\theta}|\boldsymbol{\phi})$. Suppose further that it is possible to sensibly formulate a full-rank analogue $\boldsymbol{I}_0(\boldsymbol{\theta}|\boldsymbol{\phi})$ to $\boldsymbol{I}_n(\boldsymbol{\theta}|\boldsymbol{\phi})$ that is to represent the case where n is set to its "minimum" value. For example, this quantity might be units in the average rate of growth, $\boldsymbol{I}_0(\boldsymbol{\theta}|\boldsymbol{\phi}) \approx n^{-1}\boldsymbol{I}_n(\boldsymbol{\theta}|\boldsymbol{\phi})$, or it might be devised by substituting into $\boldsymbol{I}_n(\boldsymbol{\theta}|\boldsymbol{\phi})$ the minimal sample-size information thought necessary to begin to understand the phenomenon under study. For further insight, consider that when $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ is an independent and identically distributed sample, Fisher information is $\boldsymbol{I}_n(\boldsymbol{\theta}|\boldsymbol{\phi}) = n\boldsymbol{I}_0(\boldsymbol{\theta}|\boldsymbol{\phi})$, which identifies the quantity $\boldsymbol{I}_0(\boldsymbol{\theta}|\boldsymbol{\phi})$ explicitly.

Assuming a sensible $\boldsymbol{I}_0(\boldsymbol{\theta}|\boldsymbol{\phi})$ is available, a *scaled* unit-information prior takes the form

$$\pi(\boldsymbol{\theta}|\boldsymbol{\phi}) \;=\; (2\pi\tau^2)^{-\nu/2}|\boldsymbol{I}_0(\boldsymbol{\theta}^0,\boldsymbol{\phi})|^{1/2}h(\boldsymbol{\theta}|\boldsymbol{\phi}), \tag{22}$$

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

$$h(\boldsymbol{\theta}|\boldsymbol{\phi}) \;=\; f\left(\frac{1}{2\tau^2}(\boldsymbol{\theta}-\boldsymbol{\theta}^0)^T\boldsymbol{I}_0(\boldsymbol{\theta}^0,\boldsymbol{\phi})(\boldsymbol{\theta}-\boldsymbol{\theta}^0)\right), \tag{23}$$

for some function f, which moght be chosen so that (22) is, *e.g.*, a Gaussian prior, from $f(x) = e^{-x}$, or Cauchy prior, from $f(x) = \sqrt{2/\pi}(1+2x)^{-1}$. Within this family, the unit-information prior is defined at the setting $\tau = \tilde{\tau} = 1$, 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," according to Kass and Wasserman's (1995, p. 929) characterization.

Adopting the prior (22), set $\tilde{\tau} = 1$ and apply (18) within (21) to produce an approximation to the calibrated Bayes factor (15), given by

$$NDC_{st}(\boldsymbol{Y}|\boldsymbol{\phi}) \;\approx\; \frac{|\hat{\boldsymbol{I}}_n(\hat{\boldsymbol{\theta}},\boldsymbol{\phi})|^{1/2}}{|\boldsymbol{I}_0(\boldsymbol{\theta}^0,\boldsymbol{\phi})|^{1/2}}\frac{e^{-\frac{1}{2}\|\boldsymbol{Z}(\boldsymbol{\phi})\|^2}}{h(\hat{\boldsymbol{\theta}}|\boldsymbol{\phi})}. \tag{24}$$

This is the analogue to (14) in the Gaussian means example of Section 2.2. Kass and Wasserman (1995) derive a similar approximation to the usual Bayes factor, from the setting $\tau^2 = 1$ in (22), and explore its asymptotic properties when $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^0 + O(n^{-1/2})$. This asymptotic condition on $\hat{\boldsymbol{\theta}}$ implies $h(\hat{\boldsymbol{\theta}}|\boldsymbol{\phi}) \approx f(\mathbf{0})$, and the subsequent approximation $NDC_{st}(\boldsymbol{Y}|\boldsymbol{\phi}) \approx \exp S_{st}(\boldsymbol{Y}|\boldsymbol{\phi})$ as $n \to \infty$, having defined the modified Schwarz criterion

$$S_{st}(\boldsymbol{Y}|\boldsymbol{\phi}) \;=\; -\frac{1}{2}\|\boldsymbol{Z}(\boldsymbol{\phi})\|^2 + \log\frac{|\hat{\boldsymbol{I}}_n(\hat{\boldsymbol{\theta}},\boldsymbol{\phi})|^{1/2}}{|\boldsymbol{I}_0(\boldsymbol{\theta}^0,\boldsymbol{\phi})|^{1/2}} - \log f(\mathbf{0}). \tag{25}$$

In the case where $\boldsymbol{\phi}$ is absent and $\boldsymbol{I}_n(\boldsymbol{\theta}) = n\boldsymbol{I}_0(\boldsymbol{\theta})$, the formula (25) exactly matches Kass and Wasserman's (1995) modified Schwarz criterion, in which $f(\mathbf{0})$ adjusts for a non-Gaussian prior.

In the present development of calibration concepts, asymptotic behavior as $\tau \to \infty$ is more central than asymptotic behavior as $n \to \infty$. By this perspective, it is interesting to observe that $h(\boldsymbol{\theta}|\boldsymbol{\phi}) \to f(\mathbf{0})$ as $\tau \to \infty$, hence (24) shows that, when n is large and τ is *very* large, the calibrated Bayes factor $NDC_{st}(\boldsymbol{Y}|\boldsymbol{\phi})$ is very nearly the exponentiated Schwarz criterion in (25). This property is explored in the examples of the next section.

5 Demonstrations on example data

In this section, the proposed calibration framework is demonstrated in several example data-analyses. Special attention is paid to the impact of multiplicity adjustment, sensitivity to prior scale, and connections to the Schwarz criterion.

To interpret the results, calibrated and uncalibrated Bayes factors are transformed to twice their logarithm value, so that they may be compared against the scale proposed in Kass and Raftery (1995). For example, in a comparison of M_s *vs* M_t, larger magnitudes of $2\log BF_{st}$ or $2\log NDC_{st}$ indicate stronger evidence for M_s (if positive) or M_t (if negative). The strength of evidence is categorized into "positive," "strong," and "very strong" above the thresholds 3, 6, and 10.

Calculations are implemented using various Markov Chain Monte Carlo (MCMC) techniques for posterior simulation, described in Robert and Casella (1999). 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.

5.1 Adverse events in a vaccine trial

In this first demonstration, the multiplicity properties of the calibrated setting (1) for variable selection are explored in the analysis of adverse-event data examined in Berry and Berry (2004). Also highlighted is the calibration approach laid out in Section 4.1 for non-Gaussian contexts. In order to compare with Berry and Berry's (2004) original analysis, the reanalysis developed here incorporates as much as possible the features of the original, including its elaborate hierarchical prior. In doing so, the demonstration highlights the flexibility of the proposed approach for use with complex prior formulations. To assist the reader, references are made to the pre-calibration and calibration steps listed at the end of Section 2.1. Some readers may wish to first go through Sections 5.2 and 5.3, which demonstrate the proposed concepts in simpler contexts.

The data of this example are an array of incidence-count totals from a vaccine trial that involved control and treatment groups of $n_1 = 132$ and $n_2 = 148$ subjects. The counts are of forty pre-defined "adverse event" (AE) occurrences (*e.g.*, a rash or nausea), which are uniquely grouped into eight body systems. Corresponding notation identifies pairs of triple-subscripted data, $\boldsymbol{Y}_{jk} = (Y_{1jk}, Y_{2jk})$, where k indexes AE-type $k \in K_j$ within body

system $j \in J$, and the order of pairing reflects "control" *vs* "treatment" conditions. Raw relative frequencies, $\hat{p}_{ijk} = Y_{ijk}/n_i$, and AE-type groupings into body systems are listed below in Table 1.

The data-analysis objective is to "flag" any AE-types whose occurrence-rates are greater under the vaccine treatment. Each $Y_{ijk} \sim$ binomial (n_i, p_{ijk}), independently across $i = 1, 2$ and $(j,k) \in \Omega = \{(j,k) : j \in J, k \in K_j\}$. A model M_s is characterized by three subsets: $A_{s,0}$, which collects index-pairs (j,k) such that $p_{1jk} = p_{2jk}$; $A_{s,1}$, which collects the (j,k) such that $p_{1jk} > p_{2jk}$; and, $A_{s,2}$, which collects the (j,k) such that $p_{1jk} < p_{2jk}$. The AE-types associated with the subset $A_{s,2}$ are those to be flagged.

Define $\phi_{jk} = \frac{1}{2}(\eta_{1jk} + \eta_{2jk})$ and $\theta_{jk} = \frac{1}{2}(\eta_{1jk} - \eta_{2jk})$, having set $\eta_{ijk} = \log\{p_{ijk}/(1 - p_{ijk})\}$, and collect these quantities into the parameters $\boldsymbol{\phi} = (\phi_{jk} : (j,k) \in \Omega)$ and $\boldsymbol{\theta}_s = (\theta_{jk} : (j,k) \notin A_{s,0})$, which record ν_0 nuisance and $\nu_{s,1}$ "free" target parameters of model M_s, respectively. The likelihood function of model M_s factors into components,

$$L(\boldsymbol{\theta}, \boldsymbol{\phi}; \boldsymbol{Y}) = \prod_{(j,k) \in A_{s,0}} L(0, \phi_{jk}; \boldsymbol{Y}_{jk}) \times \prod_{(j,k) \notin A_{s,0}} L(\theta_{jk}, \phi_{jk}; \boldsymbol{Y}_{jk}) \qquad (26)$$

where each component is from an exponential family,

$$L(\theta_{jk}, \phi_{jk}; \boldsymbol{Y}_{jk}) = C_\zeta(\boldsymbol{Y}_{jk}) \exp\{(Y_{1jk} - Y_{2jk})\theta_{jk} + (Y_{1jk} + Y_{2jk})\phi_{jk} - \zeta(\theta_{jk}, \phi_{jk})\},$$

where $C_\zeta(\boldsymbol{Y}_{jk}) = \binom{n_1}{Y_{1jk}}\binom{n_2}{Y_{2jk}}$ and

$$\zeta(\theta_{jk}, \phi_{jk}) = n_1 \log\left(1 + e^{\phi_{jk} + \theta_{jk}}\right) + n_2 \log\left(1 + e^{\phi_{jk} - \theta_{jk}}\right).$$

This completes the pre-calibration Step A of identifying target and nuisance parameters.

The prior is formulated hierarchically in such a way that makes use of the arrangement of AE-types within body systems. It is described in Berry and Berry (2004) as a "three-stage" prior, but it readily collapses to the following two-stage form: Each $\phi_{jk}|\tau^2_{0A}, \tau^2_{0B} \sim G\left(0, \tau^2_{0A} + \tau^2_{0B} + \tau^2\right)$ and, independently, each $\theta_{jk}|\tau^2_{1A,j}, \tau^2_{1B} \sim G\left(0, \tau^2_{1A,j} + \tau^2_{1B} + \tau^2\right)$, for hierarchical parameters τ^2, τ^2_H, τ^2_{0A}, τ^2_{0B}, $\tau^2_{1A,j}$ for $j \in J$, τ^2_{1B}. Among the hierarchical parameters, τ^2 and τ^2_H are fixed constants to be specified explicitly, while τ^2_{0A}, τ^2_{0B}, $\tau^2_{1A,j}$, and τ^2_{1B} are modeled independently such that $\tau^2_H/\tau^2_{0A} \sim \chi_\kappa$, $\tau^2_H/\tau^2_{0B} \sim \chi_\kappa$, $\tau^2_H/\tau^2_{1A,j} \sim \chi_\kappa$, and $\tau^2_H/\tau^2_{1B} \sim \chi_\kappa$ for an additional prior parameter κ. Berry and Berry set $\tau^2 = 10$, $\tau^2_H = 2$, and $\kappa = 6$; these and other settings are examined in the analysis below.

Berry and Berry incorporate a multiplicity adjustment using a variation of the beta-binomial formulation discussed in Section 3.2. The analysis explored here proceeds differently by calibrating the analysis to avoid sensitivity to the prior scale parameter and to incorporate the multiplicity adjustment specified in (1). The prior parameter τ is most influential to scale, and so will be treated as a prior scale parameter in this setup. To accommodate the remaining prior parameters, the framework developed in previous sections is extended slightly to exploit the hierarchical aspect to the prior by applying the calibration techniques conditionally given $\boldsymbol{\phi}$ and $\boldsymbol{\tau} = (\tau_{0A}^2, \tau_{0B}^2, \tau_{1A,j}^2, \tau_{1B}^2 : j \in J)$, in effect treating *both* of these quantities as nuisance parameters, even though $\boldsymbol{\tau}$ associated with the prior. This means, *e.g.*, that expert opinion would be articulated through $\rho_{st}(\boldsymbol{\phi}, \boldsymbol{\tau}) = P[M_s|\boldsymbol{\phi}, \boldsymbol{\tau}] \,/\, P[M_t|\boldsymbol{\phi}, \boldsymbol{\tau}]$ in an uncalibrated analysis, and through $\tilde{\rho}_{st}(\boldsymbol{\phi}, \boldsymbol{\tau}) = P[M_s|\tilde{\boldsymbol{Y}},\ \boldsymbol{\phi},\ \boldsymbol{\tau}] \,/\, P[M_t|\tilde{\boldsymbol{Y}}, \boldsymbol{\phi}, \boldsymbol{\tau}]$ in a calibrated analysis. Note the convenience of additionally conditioning on $\boldsymbol{\tau}$, both for conceptual formulation and calculation, due to the availability of a closed form for $\pi(\boldsymbol{\phi}|\boldsymbol{\tau})$, and the absence of one for $\pi(\boldsymbol{\phi})$. This completes the pre-calibration Steps B and C of identifying the scale parameter, and deriving a suitable mathematical framework for working with the model conditionally, which is achieved here by treating $\boldsymbol{\tau}$ as a special type of nuisance parameter.

It is readily checked that the likelihood function and prior satisfy the conditions of Section 4.1. Referring to the terms of Theorem 2, define "sample size" as $n = n_1 + n_2$, and observe that the likelihood function implies

$$\hat{\boldsymbol{I}}_n(\boldsymbol{\theta}|\boldsymbol{\phi}) \quad = \quad n \left\{ \frac{e^{\phi_{jk}}}{1+e^{\phi_{jk}}} \left(1 - \frac{e^{\phi_{jk}}}{1+e^{\phi_{jk}}} \right) \right\}, \tag{27}$$

which increases without bound as n grows. Hence, the conclusion of Theorem 2 holds, and the target value (12) is valid for calibration. In the absence of well-studied default settings for priors of this form, the default prior concept applied here is an appeal to simplicity, setting $\tilde{\tau} = 1$ or $\tilde{\tau} = k_s$ depending on whether the multiplicity adjustment (1) is incorporated in the calibration. Given that the hierarchical parameters in $\boldsymbol{\tau}$ are negligible when τ is large, the setting $\tilde{\tau} = 1$ roughly incorporates the intuition underlying a unit-information prior.

This completes the calibration Step D of setting $\tilde{\tau}$ from a selected default-prior concept. To be clear, two separate calibrations are examined in this demonstration, one derived from the "unit-information" concept, for which $\tilde{\tau} = 1$, and the other that adopts the default setting implied in Theorem 1, for which $\tilde{\tau} = k_s$, to induce a multiplicity adjustment. The appeal to

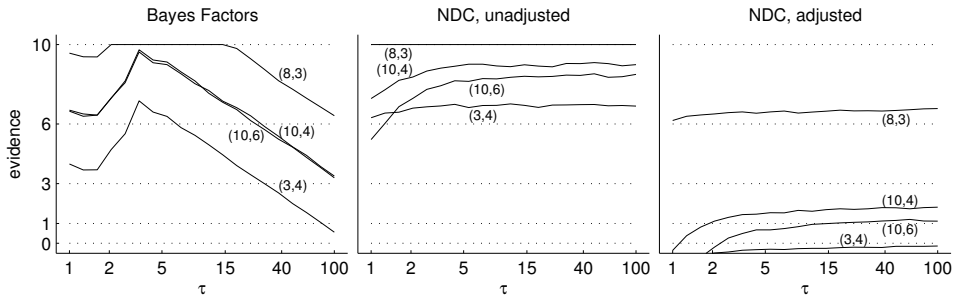

Figure 1 Evidence assessments of an adverse event on Berry and Berry's (2004) data for τ between 1 and 100, plotted on a standard scale of evidence. The left panel plots transformed Bayes factors, the middle panel plots transformed calibrated Bayes factors that are unadjusted for multiplicity, and the right panel plots transformed calibrated Bayes factors that are adjusted for multiplicity. Assessments are reported only of the four AE-types indexed by $(j, k) = (3, 4)$, $(8, 3)$, $(10, 4)$, and $(10, 6)$.

Theorem 2 completes the calibration Steps E and F. By that theorem, the corresponding values of $BF_{st}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi}, \boldsymbol{\tau})$ are directly ascertained as the limiting values $\tau/\tilde{\tau}$ identified in formula (20), having implicitly solved for the anchor point $\tilde{\boldsymbol{Y}}$.

Calculations are made using the reversible-jump MCMC algorithm, within a Gibbs structure to integrate across parameters. The algorithm is implemented by making reversible jumps on individual θ_{jk} at fixed values of the remaining parameters. Proposed jumps from $(\theta^0_{s,jk}, \phi_{s,jk})$ in model M_s to $(\theta_{t,jk}, \phi_{t,jk})$ in model M_t are defined through the invertible transformation $\phi_{t,jk} = \frac{1}{4}\phi_{s,jk}\{4 + a(\boldsymbol{u})\}$ and $\theta_{t,jk} = \frac{1}{4}\phi_{s,jk}\{4 - a(\boldsymbol{u})\}$, where $a(\boldsymbol{u}) = 1/\boldsymbol{u} - 1/(1 - \boldsymbol{u})$ and $\boldsymbol{u} \sim \text{Beta}(2, 2)$. Acceptance probabilities are calculated using the formula (7) for calibrated prior odds, extended slightly to $\rho_{st}(\boldsymbol{\phi}, \boldsymbol{\tau}) = \tilde{\rho}_{st}(\boldsymbol{\phi}, \boldsymbol{\tau})/BF_{st}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi}, \boldsymbol{\tau})$ so as to additionally to condition upon $\boldsymbol{\tau}$.

Results of the analysis, in a variety of configurations, are indicated in Figure 1 and Table 1. Data analysis is carried out repeatedly across twenty values of the scale parameter τ in the range $1 \leq \tau \leq 100$, which forms the horizontal axis in each panel of Figure 1, all while holding constant the ratio $\tau^2/\tau^2_H = 5$, and the parameter $\kappa = 6$. Each panel plots the relevant assessments after having been transformed according to $2\log\{P[\mathcal{E}_{(j,k)}|\boldsymbol{Y}]/(1 - P[\mathcal{E}_{(j,k)}|\boldsymbol{Y}])\}$, where $\mathcal{E}_{(j,k)}$ collects all models M_s such that $(j, k) \in A_{s,2}$. That is, the values plotted in Figure 1 indicate support for an adverse event of type j in body-

				B&B		NDC, unadj.		NDC, adj.	
j	k	$\hat{p}_{1jk}$	$\hat{p}_{2jk}$	eq.	AE	eq.	AE	eq.	AE
1	1	0.303	0.385	0.762	0.211	0.392	0.601	0.956	0.043
1	2	0.197	0.230	0.827	0.122	0.798	0.180	0.992	0.007
1	3	0.000	0.014	0.796	0.101	0.054	0.944	0.947	0.053
1	4	0.008	0.020	0.813	0.100	0.765	0.216	0.992	0.008
1	5	0.152	0.182	0.826	0.116	0.798	0.179	0.993	0.007
3	1	0.015	0.047	0.821	0.117	0.328	0.666	0.949	0.050
3	2	0.000	0.014	0.835	0.083	0.067	0.932	0.738	0.261
3	3	0.000	0.014	0.812	0.101	0.068	0.930	0.859	0.141
3	**4**	**0.076**	**0.162**	**0.743**	**0.231**	**0.030**	**0.969**	**0.517**	**0.483**
3	5	0.008	0.020	0.823	0.093	0.767	0.214	0.992	0.008
3	6	0.053	0.014	0.805	0.050	0.211	0.003	0.910	0.000
3	7	0.144	0.128	0.849	0.076	0.890	0.044	0.996	0.002
5	1	0.015	0.020	0.717	0.136	0.882	0.083	0.996	0.003
6	1	0.015	0.000	0.666	0.087	0.039	0.001	0.431	0.000
8	1	0.000	0.014	0.655	0.185	0.023	0.977	0.749	0.251
8	2	0.015	0.014	0.661	0.153	0.898	0.048	0.997	0.001
8	**3**	**0.326**	**0.507**	**0.214**	**0.780**	**0.001**	**0.999**	**0.033**	**0.967**
9	1	0.008	0.027	0.900	0.059	0.560	0.428	0.981	0.019
9	2	0.015	0.027	0.901	0.058	0.828	0.147	0.994	0.005
9	3	0.015	0.007	0.896	0.040	0.858	0.026	0.995	0.001
9	4	0.061	0.088	0.906	0.062	0.758	0.223	0.991	0.008
9	5	0.152	0.189	0.897	0.083	0.754	0.228	0.990	0.009
9	6	0.008	0.014	0.898	0.047	0.870	0.101	0.996	0.003
9	7	0.061	0.088	0.906	0.061	0.758	0.223	0.991	0.008
9	8	0.106	0.101	0.904	0.051	0.894	0.055	0.997	0.002
9	9	0.008	0.020	0.903	0.051	0.766	0.215	0.992	0.008
9	10	0.008	0.014	0.905	0.042	0.870	0.101	0.996	0.003
9	11	0.008	0.020	0.907	0.050	0.769	0.212	0.992	0.008
10	1	0.000	0.027	0.859	0.087	0.001	0.999	0.065	0.935
10	2	0.000	0.014	0.860	0.070	0.001	0.999	0.945	0.054
10	3	0.008	0.014	0.868	0.062	0.872	0.099	0.996	0.003
10	**4**	**0.023**	**0.088**	**0.784**	**0.190**	**0.011**	**0.989**	**0.287**	**0.713**
10	5	0.015	0.041	0.852	0.099	0.540	0.450	0.978	0.022
10	**6**	**0.008**	**0.054**	**0.836**	**0.126**	**0.014**	**0.986**	**0.364**	**0.636**
10	7	0.015	0.027	0.862	0.076	0.828	0.148	0.994	0.005
10	8	0.015	0.000	0.852	0.048	0.009	0.000	0.798	0.000
10	9	0.015	0.007	0.855	0.055	0.857	0.026	0.995	0.001
11	1	0.015	0.000	0.721	0.079	0.008	0.000	0.789	0.000
11	2	0.106	0.122	0.757	0.102	0.858	0.111	0.995	0.004
11	3	0.008	0.014	0.749	0.121	0.872	0.099	0.996	0.003

Table 1 *Posterior probabilities on Berry and Berry's (2004) adverse-event data. Index values are listed for body system (j) and AE-type (k) in the first pair of columns; the remaining columns list raw adverse-event relative frequencies, followed by the posterior probabilities from Berry and Berry's (2004) analysis (headed "B & B"), then those derived from calibrated Bayes factors unadjusted for multiplicity, and finally those derived from calibrated Bayes factors that are adjusted for multiplicity. The columns labeled "eq." list posterior probabilities that the rates of adverse events between treatment and control conditions are equal, and those labeled "AE" list posterior probabilities of an adverse event.*

system k, transformed for interpretation on Kass and Raftery's (1995) scale of evidence. Only four AE-types are represented in Figure 1, those with index values $(j,k) = (3,4)$, $(8,3)$, $(10,4)$, and $(10,6)$, which were selected by Berry and Berry for having been flagged in a previous frequentist analysis. Selected results for all AE-types are listed in Table 1, but only at $\tau = 100$.

The three panels of Figure 1 show results of the proposed procedure in three configurations of the discrete prior: the configuration represented in the left panel has $\rho_{st}(\boldsymbol{\phi}, \boldsymbol{\tau}) = 1$, hence the results shown are from Bayes factors; that of the middle panel has $\tilde{\rho}_{st}(\boldsymbol{\phi}, \boldsymbol{\tau}) = 1$ and $\tilde{\tau} = 1$, hence the results are from calibrated Bayes factors that are calibrated to avoid sensitivity to the prior, but do not incorporate a multiplicity adjustment; and, that of the right panel has $\tilde{\rho}_{st}(\boldsymbol{\phi}, \boldsymbol{\tau}) = 1$ and $\tilde{\tau} = k_s$, which is as in the middle panel but with a multiplicity adjustment. The latter two configurations are also represented in Table 1, as posterior probabilities, alongside Berry and Berry's results for comparison.

As expected, and illustrated in Figure 1, support for an adverse event drastically weakens as τ grows large when it is reported as a Bayes factor, but it eventually stabilizes when it is reported as a calibrated Bayes factor. "Strong" evidence of an AE (a reported value above 6) of every selected type is indicated in the middle panel, with each calibrated Bayes factor stabilizing (by coincidence) near the maximum of the corresponding Bayes factor in the left panel. Comparison with the right panel illustrates how the multiple-model adjustment weakens evidence across the board, so much that "strong" support of an AE remains only for the AE-type at $(j,k) = (8,3)$. The adjustment ultimately induces a beneficial clarifying effect of reducing the collection of several suspicious AE-types to just one that is to be flagged.

In Table 1, it is seen that Berry and Berry's prior weakens the reported evidence even more, to the point where no strong evidence of an AE is exhibited among any of the forty AE-types. Consider that on Berry and Berry's results the transformation $2\log\{P[\mathcal{E}_{(j,k)}|\boldsymbol{Y}]/(1-P[\mathcal{E}_{(j,k)}|\boldsymbol{Y}])\}$ yields the values -2.41, 2.53, -2.90, and -3.87 for $(j,k) = (3,4)$, $(8,3)$, $(10,4)$, and $(10,6)$. These transformed assessments are much different than those of the multiplicity-adjusted calibrated Bayes factors, and it is interesting that Berry and Berry's results are also hard to place among the patterns exhibited in Figure 1: even at $\tau = 100$, the Bayes factors in the left panel report much stronger evidence than those of Berry and Berry, and yet the calibrated Bayes factors of the other two panels are well past the point of having stabilized with respect to τ. From this perspective, the configuration introduced in Berry and Berry's hierarchical discrete prior is seen to have

an astoundingly strong effect.

5.2 The Behrens-Fisher problem

This next example demonstrates the proposed methodology in the context of the Behrens-Fisher problem, a setup that has been studied by many authors, frequentist and Bayesian. It is important in forensic "matching" applications in which measurements of trace material (*e.g.*, glass fragments) or pattern marks (*e.g.*, fingerprints) found at a crime scene are compared to those on a suspect; the aim is to quantify the strength of evidence that the two sets of measurements are from the same source. See *e.g.*, Lindley (1977) and Lund and Iyer (2017) for further discussion of such applications. The present exploration uses data from a simpler application, the "yarn strength" data from Box and Tiao (1992, ex. 2.5.4).

The Behrens-Fisher 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. The example 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 M_0 puts $\boldsymbol{Y}_i|\mu, \sigma_i^2 \sim G(\mu\mathbf{1}, \sigma_i^2\boldsymbol{I}_{n_i})$ and M_1, puts $\boldsymbol{Y}_i|\mu_i, \sigma_i^2 \sim G(\mu_i\mathbf{1}, \sigma_i^2\boldsymbol{I}_{n_i})$.

Although formula (22) defines a scaled unit-information prior from Fisher information, it is possible in this problem to identify such a prior from direct arguments. Observe that, under M_1,

$$\frac{n_1\bar{Y}_1/\sigma_1^2 + n_2\bar{Y}_2/\sigma_2^2}{n_1/\sigma_1^2 + n_2/\sigma_2^2}\,\Bigg|\,\sigma_1^2, \sigma_2^2 \sim G\left(\frac{n_1\mu_1/\sigma_1^2 + n_2\mu_2/\sigma_2^2}{n_1/\sigma_1^2 + n_2/\sigma_2^2}, \frac{1}{n_1/\sigma_1^2 + n_2/\sigma_2^2}\right) \quad (28)$$

and, independently,

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sigma_1^2/n_1 + \sigma_2^2/n_2}\,\Bigg|\,\sigma_1^2, \sigma_2^2 \sim G\left(\frac{\mu_1 - \mu_2}{\sigma_1^2/n_1 + \sigma_2^2/n_2}, \frac{1}{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right). \quad (29)$$

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 concept to (28) and (29) motivates the transformation

$$\mu = \frac{\mu_1/\sigma_1^2 + \mu_2/\sigma_2^2}{1/\sigma_1^2 + 1/\sigma_2^2} \quad \text{and} \quad \theta = \frac{\mu_1 - \mu_2}{\sigma_1^2 + \sigma_2^2}, \quad (30)$$

and identifies the corresponding conditional scaled unit-information priors

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

where τ is the scale parameter. The variance parameters are assigned independent scaled inverse-chi-square distributions, $\lambda/\sigma_1^2 \sim \chi_\kappa^2$ and $\lambda/\sigma_2^2 \sim \chi_\kappa^2$, as priors.

Under the transformation (30), the model 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 M_0 is specified by the setting $\theta=\theta_0=0$. The relevant conditional Bayes factor is

$$BF_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) \quad = \quad \left(1+\tau^2\frac{n_1\sigma_1^2+n_2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^{1/2} \exp\left\{-\frac{1}{2}w(\sigma_1^2,\sigma_2^2)Z(\boldsymbol{\phi})^2\right\} \quad (31)$$

where

$$Z(\boldsymbol{\phi})^2 = \frac{\{n_1(\bar{Y}_1-\mu)-n_2(\bar{Y}_2-\mu)\}^2}{n_1\sigma_1^2+n_2\sigma_2^2}$$

and

$$w(\sigma_1^2,\sigma_2^2) = \frac{\tau^2(n_1\sigma_1^2+n_2\sigma_2^2)/(\sigma_1^2+\sigma_2^2)}{1+\tau^2(n_1\sigma_1^2+n_2\sigma_2^2)/(\sigma_1^2+\sigma_2^2)}.$$

This completes the pre-calibration steps listed at the end of Section 2.1. The derived framework for conditioning allows the subsequent calibration steps to be implemented using the arguments put forward in Section 2.2 for the Gaussian means problem: by adopting the default value $\tilde{\tau}=1$, which is implied from the unit-information default-prior concept, and incorporating the target value (12), the calibrated Bayes factor is

$$NDC_{01}(\boldsymbol{Y}|\boldsymbol{\phi}) = \left(\frac{1}{\tau^2}+\frac{n_1\sigma_1^2+n_2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)^{1/2} \exp\left\{-\frac{1}{2}w(\sigma_1^2,\sigma_2^2)Z(\boldsymbol{\phi})^2\right\}. \quad (32)$$

The present exploration also considers a second version of the analysis in which the Gaussian prior on the target parameter is replaced with a Cauchy prior,

$$\theta|\sigma_1^2,\sigma_2^2 \quad \sim \quad \text{Cauchy}\left(0, \tau\sqrt{\frac{1}{\sigma_1^2+\sigma_2^2}}\right).$$

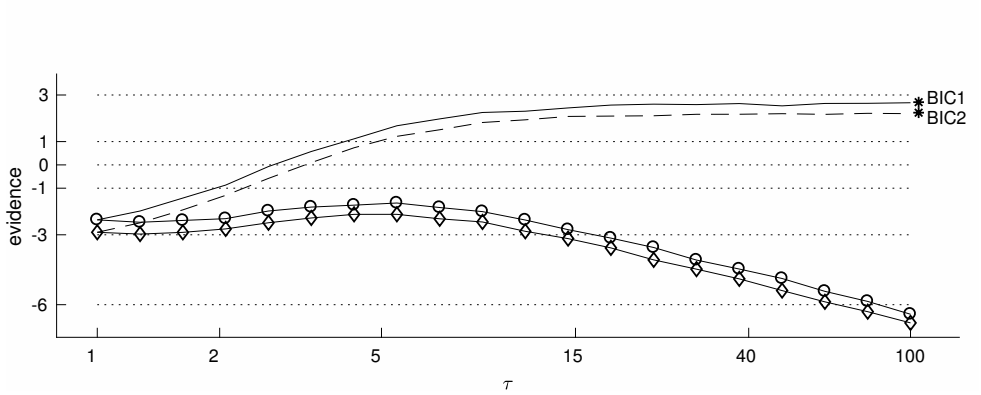

Figure 2 Evidence assessments on Box and Tiao's yarn-strength data for τ between 1 and 100, plotted on a standard scale of evidence. The solid line and dashed lines plot evidence from calibrated Bayes factors under a Gaussian and Cauchy prior, respectively. The solid lines overlaid with circles or diamonds plot evidence from uncalibrated Bayes factors under a Gaussian and Cauchy prior, respectively. The points BIC1 and BIC2 mark values of the Schwarz criterion, using "MLE substitution" to handle nuisance parameters, under a Gaussian and Cauchy prior, respectively.

This is consistent with a recommendation of Jeffreys (1961), by which the default setting is $\tilde{\tau} = 1$. A Cauchy prior is also recommended in Liang *et al.* (2008) in order to avoid the "information paradox," a phenomenon that occurs in Gaussian testing problems when variance parameters are estimated.

The calculation of unconditional versions of a calibrated or uncalibrated Bayes factor is implemented using the formula

$$P[M_0|\boldsymbol{Y}] \;=\; \left\{1 + \frac{\int P[M_1|\boldsymbol{Y},\boldsymbol{\phi}]\pi_0(\boldsymbol{\phi}|\boldsymbol{Y})d\boldsymbol{\phi}}{\int P[M_0|\boldsymbol{Y},\boldsymbol{\phi}]\pi_1(\boldsymbol{\phi}|\boldsymbol{Y})d\boldsymbol{\phi}}\right\}^{-1}, \tag{33}$$

which converts conditional posterior model probabilities to unconditional posterior model probabilities. Here, $\pi_0(\boldsymbol{\phi}|\boldsymbol{Y})$ and $\pi_1(\boldsymbol{\phi}|\boldsymbol{Y})$ denote the model-specific posterior densities of the nuisance parameter. Integration in (33) is carried out numerically by averaging over model-specific MCMC-generated samples.

Figure 2 displays unconditional assessments calculated from calibrated and uncalibrated Bayes factors, under both the Gaussian and Cauchy specifications of the prior. Twenty values of the scale parameter are examined, across the range $1 \leq \tau \leq 100$, which indexes the horizontal axis of Figure 2. The prior variance parameters are set to $\kappa = \lambda = 0$ in every evaluation, a standard setting that is applied for illustration. The precise quantities that are plotted in Figure 2 are manifestations of the formula

$2\log(P[M_1|\boldsymbol{Y}]/P[M_0|\boldsymbol{Y}])$, by which larger magnitudes indicate stronger evidence for M_1.

In Figure 2, the solid lines overlaid with circles or diamonds are calculated from the uncalibrated Bayes factor (31), under a Gaussian or Cauchy prior, respectively. The solid and dashed lines calculated from the calibrated Bayes factor (32), under a Gaussian or Cauchy prior, respectively. As expected from its scaling properties, the evidence for M_1 exhibited by the Bayes factor grows drastically weaker as τ increases beyond a certain value, while that exhibited by the calibrated Bayes factor eventually stabilize. The results are also consistent in illustrating that evidence for M_1 is weaker under a Cauchy prior.

Several assessments alluded to in previous discussion, but not addressed in detail, are also plotted in Figure 2. 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 values are calculated by evaluating the formula (25) and applying "MLE substitution" to handle the nuisance parameters. Refer to Bollen *et al.* (2012) for general discussion of such techniques. The formula is

$$\hat{S}_{01}(\boldsymbol{Y}) \;=\; \frac{1}{2}Z(\hat{\boldsymbol{\phi}})^2 - \frac{1}{2}\log\left(\frac{n_1\hat{\sigma}_1^2 + n_2\hat{\sigma}_2^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}\right) - \log f(\mathbf{0}),$$

where $\hat{\boldsymbol{\phi}} = (\hat{\mu}, \hat{\sigma}_1^2, \hat{\sigma}_2^2)$ is the maximum-likelihood value of $\boldsymbol{\phi}$ under M_1. The value BIC1 is calculated using $f(x) = e^{-x}$, which is associated with a Gaussian prior, and BIC2 is calculated using $f(x) = \sqrt{2/\pi}(1+2x)^{-1}$, which is associated with a Cauchy prior. It is reassuring that the results derived from the calibrated Bayes factors tend to stabilize near those of the *ad hoc* Schwarz criteria. It will be seen in the example analysis of Section 5.2, below, that this pattern can be disrupted in more complicated testing contexts.

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

The example of this section 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

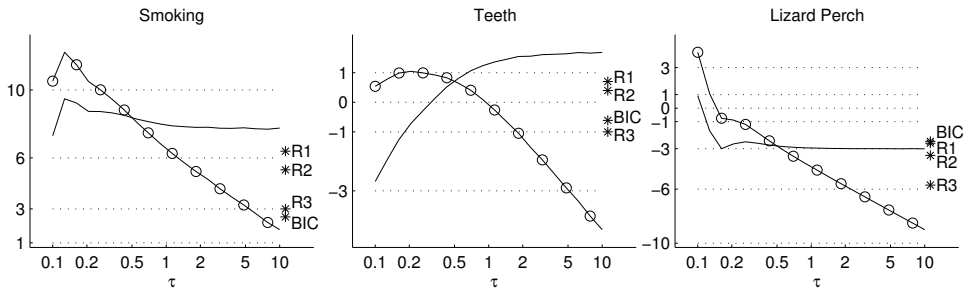

Figure 3 Evidence assessments on Raftery's "Smoking," "Teeth," and "Lizard Perch" tables for τ between 0.1 and 10, plotted on a standard scale of evidence. The solid line marks calibrated Bayes factors; the solid line with circles marks Bayes factors. The asterisks labelled "R1," "R2," "R3," mark the assessments obtained under Raftery's prior at respective values 1, 1.65, and 5 of that prior's scale parameter.

the "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 M_0 and M_1 are distinguished by the absence, in M_0, or presence, in 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 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 target parameter $\theta = (\eta_{11} - \eta_{12} - \eta_{21} + \eta_{22})/2$ identifies the magnitude of interaction; it is fixed at $\theta = \theta_0 = 0$ under model 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.

This completes the pre-calibration Steps A and C listed at the end of Section 2.1. A scaled unit-information prior is to be specified, but its formulation will be clearer upon first confirming the assumptions of Theorem 2, which will incidentally ready the completion of the calibration Steps E and F by prescribing that $BF_{st}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi})$ be set to the limiting value $\tau/\tilde{\tau}$ identified in formula (20).

To work with Theorem 2, a suitable asymptotic framework is needed. For that, treat $n(\theta_0, \boldsymbol{\phi})$ as "sample size," and assume that each $E[Y_{jk}] = e^{\eta_{jk}}$ is asymptotically similar to $n(\theta_0, \boldsymbol{\phi})$ as the latter quantity becomes arbitrarily large, *i.e.*, the ratios of one to the other of these quantities are both bounded. 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 (18), and is otherwise consistent with the framework of Section 4. 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 M_0.

To formulate a scaled unit-information prior, start by making the straightforward deduction 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. Accordingly, the scaled unit-information prior adopted 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 M_1, and a constrained version of the same prior under model M_0. Having adopted a scaled unit-information prior, which identifies the scale parameter and its default value as part of its formulation, this completes the remaining pre-calibration Step B and calibration Step D listed at the end of Section 2.1.

Analysis results on Raftery's count data, calculated at several settings of τ^2, are plotted in Figure 3. As in previous example analyses, the scale parameter is examined across of range of twenty values, this time across $0.1 \leq \tau \leq 10$, which form the horizontal axis of each panel; as before, the quantities plotted are $2\log(P[M_1|\boldsymbol{Y}]/P[M_0|\boldsymbol{Y}])$, calculated from a Bayes factor, with $\rho_{01}(\boldsymbol{\phi}) = 1$, or calibrated Bayes factor, with $\tilde{\tau} = 1$ and $\tilde{\rho}_{01}(\boldsymbol{\phi}) = 1$, either of which indicate the strength of evidence for the model M_1. As in the example of Section 5.2, computations rely on MCMC simulation, together with formula (33). In each panel of Figure 3, one sees the same pattern observed previously, wherein the uncalibrated Bayes factor exhibits increasingly stronger evidence for M_0 at larger values of τ, while the calibrated Bayes factor stabilizes.

For reference, results associated with the Bayes factors calculated in Raftery (1993) are marked in each panel of Figure 3 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 calibrated Bayes factors 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 calibrated Bayes factors for large τ^2; in the other examples, the strength of evidence indicated by BIC for M_1, relative to calibrated Bayes factors, is substantially weaker.

6 Conclusions

The goal of this article is to develop testing methodology for Bayesian practice that promotes the voice of the expert in scenarios where it is desirable that statistical procedures exhibit certain analytically-derived properties. The proposed scheme is fully coherent, and does not deviate from the mathematical formulation of subjective Bayesian analysis. It offers the expert flexibility to specify the continuous portion of the prior using intuition from estimation, and the option to retain influence on the discrete portion of the prior and the assessment of evidence through the quantities $\tilde{\rho}_{st}(\boldsymbol{\phi})$. The methodology is proposed with the hope of cultivating greater involvement by experts in specifying prior knowledge.

Its goal is achieved by reducing the role of default priors from outright substitutes for prior knowledge to tools used to calibrate Bayes factors. A calibration scheme is proposed that operates by anchoring analysis to a point in the data space, under the guidance of a default-prior concept. An unambiguous rule (19) is provided for defining the anchor point in a meaningful way, in the sense that it may be interpreted in terms of a conceptual exercise of imagining "neutral" data. Theoretical results establish that the scheme is relevant and practical to apply within a widely-used class of regular models. Its feasibility is supported by demonstrations on example data.

It is worthwhile to delineate between the subjective and objective Bayes aspects of the proposed calibration scheme. Observe that the continuous

portion of the prior is determined before the data are observed or the likelihood function is formulated, and in that aspect the scheme is entirely subjectivist. This points to one of the scheme's major practical benefits, which is that the analyst need not adopt a different mindset between estimation and testing, but may use the same continuous prior for both. The discrete portion of the prior is defined through calibrated prior odds, $\rho_{st}(\phi)$, as in (17), which is a combination of expert knowledge and a calibration rule that builds on a chosen default-prior concept. The calibrated prior odds is determined before the data are collected (*i.e.*, there is no "double-use" of data), and defines a model that is a plausible result of expert elicitation. Nevertheless, because default-prior concepts are typically defined from the likelihood function, there is reason to regard discrete prior's subjectivism as having been compromised.

In particular, from the viewpoint of personal probability, it may be unclear *whose* prior defines the quantity $\rho_{st}(\phi)$, despite that its uncalibrated analogue, $\tilde{\rho}_{st}(\phi)$, is the expression of an expert. One the other hand, it is not difficult to envision an alternative use for the ideas presented here as tools for guiding the back-and-forth between expert and analyst in an actual elicitation process, wherein the expert, having through this process gained some understanding of the nuances of Bayesian testing, ultimately becomes invested in the value identified for $\rho_{st}(\phi)$. Surely such alternative use requires development, which is not attempted here, but it points to a setting in which the proposed methodology would be entirely subjective in character.

The article furthermore explores a particular multiplicity adjustment for variable selection problems that is related to the truncated Poisson prior of Womack *et al.* (2015). In the present exploration, its asymptotic consistency properties are established under weak dimensionality assumptions. It is reformulated as a calibration, and demonstrated in an analysis of an example data set of adverse-event responses, where it is shown to have a strong clarifying effect for flagging worrisome adverse-event types.

The proposed methodology is also examined for its connection to Schwarz's (1978) model-choice criterion. The exploration of Section 4.2 suggests an interpretation by which a scaled unit-information prior (22) places a calibrated Bayes factor on a spectrum falling between a default-configured Bayes factor and the exponentiated Schwarz criterion, highlighting that a calibrated assessment of evidence is robust to modifications of the prior, but it is still sensitive to expert opinion. This property is reflected in the example analyses of Section 5; however, those demonstrations also show that the connection to the Schwarz criterion may be disrupted when *ad hoc* adjustments are

incorporated in the Schwarz criterion to account for nuisance parameters.

The proposed methodology makes heavy use of conditioning in order to deal with nuisance parameters. The example analysis of log-linear models in Section 5.2 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 $n(\theta_0, \boldsymbol{\phi})$ for that analysis. The possibilities for meeting the requirements of intuition are widened by the added flexibility of parameter-specific formulations for sample size.

Although this article focuses on multiplicity adjustments and prior sensitivity, an expectation is stated in Section 1 that the proposed calibration framework has broader applicability. As one example of a potential application not explicitly discussed here, the reader is invited to ponder methodology for high-dimensional global testing under a smoothness assumption, as would be applied in non-parametric regression or functional data analysis. For a canonical version of those contexts, explored in Spitzner (2008), consider working with "preprocessed" data that are either the Fourier coefficients of a smooth function measured densely, with error, or the average of Fourier coefficients of a random sample of densely measured smooth functions. For either situation, suppose the coefficients are collected into a high-dimensional summary vector of p independent components, $\boldsymbol{Y} = (\bar{Y}_1, \ldots, \bar{Y}_p)$, in which each component is modeled as $\bar{Y}_i|\theta_i \sim G(\theta_i, 1/n)$. The "global" test is of the model-comparison M_0 *vs* M_1, where M_0 has all $\theta_i = 0$, and M_1 treats all θ_i as free parameters. A suitable prior for M_1 has $\theta_i \sim G(0, \tau_i^2)$, independently across i. Smoothness may be understood through such concepts as a "Sobolev" geometry, which motivates a configuration of the prior such that its scale parameters taper to zero, $\tau_i \to 0$. Moreover, it is possible to set the tapering rate in such a way as to achieve a desirable analytical property. Should expert opinion specify a different rate, or no tapering at all, then the calibration techniques developed here would be useful for managing the discrepancy.

Appendix A: Appendix

Proof. (Theorem 1) It is equivalent to establish the boundedness of $R_s = 1/P[M_s|\boldsymbol{Y}]$. Writing this quantity as a sum of non-negative terms, $R_s = \sum_{t\in S}\{P[M_t|\boldsymbol{Y}]/P[M_s|\boldsymbol{Y}]\}$, as is implied by $\sum_{t\in S} P[M_t|\boldsymbol{Y}] = 1$, admits use of an extension of the Borel-Cantelli lemma (*cf.* Billingsley, 1995, prob. 22.3, p. 294), which provides that the boundedness of $R_s = 1/P[M_s|\boldsymbol{Y}]$ follows from that of $E[R_s|\boldsymbol{\theta}]$.

Substitution of each calibrated Bayes factor in (18) with $NDC_{uv}(\boldsymbol{Y}) = k_u BF_{uv}(\boldsymbol{Y})/\tau$ yields

$$\frac{P[M_s|\boldsymbol{Y}]}{P[M_t|\boldsymbol{Y}]} \;=\; \left\{\frac{k_s!}{(k_s-j_1)!}\frac{BF_{su_{j_1}}(\boldsymbol{Y})}{\tau^{j_1}}\right\}\left\{\frac{(k_s-j_1)!}{(k_s-j_1+j_2)!}\frac{\tau^{j_2}}{BF_{tu_{j_1}}(\boldsymbol{Y})}\right\}.$$

For M_u *vs* M_v such that $A_u - A_v = \{i\}$, the Bayes factor is $BF_{uv}(\boldsymbol{Y}) = (1+\tau^2 n)^{1/2}\exp\{-\frac{1}{2}w_n Z_i^2\}$, where $Z_i = n^{1/2}\bar{Y}_i$, $\bar{Y} = n^{-1}\sum_{j=1}^{n} Y_{ij}$, and $w_n = \tau^2 n/(1+\tau^2 n)$. Conditional on θ_i, the quantity Z_i^2 is a non-central chi-square random variable with one degree of freedom and non-centrality parameter $n\theta_i^2$. It follows from the chi-square moment-generating function formula, $M(t) = (1-2t)^{-1/2}\exp\{n\theta_i^2 t/(1-2t)\}$, that for $r = 1,\ldots,j_1$, since $i \in A_s$ has $\theta_i = 0$, $E\left[1/BF_{u_{r-1}u_r}(\boldsymbol{Y})\middle|\boldsymbol{\theta}_s\right] = 1$; and, for $r = j_1+1,\ldots,j_1+j_2$, since $i \notin A_s$ has $\xi_n \le |\theta_i|$, $E\left[BF_{u_r u_{r-1}}(\boldsymbol{Y})\middle|\boldsymbol{\theta}_s\right] \le \tau\zeta_n$, where

$$\zeta_n \;=\; \left(\frac{1+\tau^2 n}{\tau^2}\right)^{1/2}\left(\frac{1+\tau^2 n}{1+2\tau^2 n}\right)^{1/2}\exp\left\{-\frac{1}{2}\left(\frac{\tau^2 n}{1+2\tau^2 n}\right)n\xi_n^2\right\}.$$

The above insights set up a partition of the index space, S, relative to $s \in S$, into subsets corresponding to the unique pairs (j_1, j_2), for $j_1 = 0,\ldots,k_s$ and $j_2 = 0,\ldots,\nu_s$ such that the subset associated with a given pair (j_1, j_2) indexes the $\binom{k_s}{j_1}\binom{\nu_s}{j_2}$ model-configurations that each selects j_1 variables among those omitted in M_s and omits j_2 variables among those selected in M_s. It follows that

$$E[R_s|\boldsymbol{\theta}_s] \;\le\; R_s^* = \sum_{j_1=0}^{k_s}\sum_{j_2=0}^{\nu_s}\binom{k_s}{j_1}\binom{\nu_s}{j_2}\frac{(k_s-j_1)!}{k_s!}\tau^{j_1}\frac{(k_s-j_1+j_2)!}{(k_s-j_1)!}\zeta_n^{j_2}.$$

This is evaluated by separately considering the "outer sum" and "inner sum" operations of the double sum, $R_s^* = \sum_{j_1=0}^{k_s} U_{j_1}^{out}\sum_{j_2=0}^{\nu_s} U_{j_2}^{in}(j_1)$, where

$$U_{j_1}^{out} = \binom{k_s}{j_1}\frac{(k_s-j_1)!}{k_s!}\tau^{j_1} \quad\text{and}\quad U_{j_2}^{in}(j_1) = \binom{\nu_s}{j_2}\frac{(k_s-j_1+j_2)!}{(k_s-j_1)!}\zeta_n^{j_2}.$$

Observe that at $j_2 = 0$ one has $U_{j_2}^{in}(j_1)$. Otherwise, if $1 \le j_2 \le \nu_s$, use Sterling's approximation, $\sqrt{2\pi}k^{k+1/2}e^{-k} \le k! \le ek^{k+1/2}e^{-k}$, to see that

$$\begin{aligned} U_{j_2}^{in}(j_1) \le\; & B_0 B_1(j_2)^{k_s-j_1+1/2}B_3(j_2)^{\nu_s+1/2} \\ & \times\frac{1}{j_2!}\exp\left[j_2\left\{\log\zeta_n + \log B_2(j_2) + \log B_4(j_2) - 2\right\}\right] \end{aligned} \tag{34}$$

where B_0 is a fixed constant, $B_1(j_2) = 1 + j_2/(k_s - j_1)$ if $j_1 < k_s$ and $B_1(j_2) = j_2$ if $j_1 = k_s$; $B_2(j_2) = k_s - j_1 + j_2$ if $j_2 > 0$ and $B_2(j_2) = 1$ if $j_2 = 0$; $B_3(j_2) = 1 + j_2/(\nu_s - j_2)$ if $j_2 < \nu_s$ and $B_3(j_2) = j_2$ if $j_2 = \nu_s$; and, $B_4(j_2) = \nu_s - j_2$ if $j_2 < \nu_s$ and $B_4(j_2) = 1$ if $j_2 = \nu_s$. The exponent in the final factor of (34) is bounded above by $j_2(\log \zeta_n + 2 \log p_n - 2)$. It follows, after moving initial terms into the exponent, that

$$U^{in}_{j_2}(j_1) \leq B_0 \times \frac{1}{j_2!} \exp\left[j_2 \left\{\log \zeta_n + 2 \log p_n - 2 + g(j_2)\right\}\right] \tag{35}$$

where

$$g(j_2) \;\; = \;\; j_2^{-1} \left\{(k_s - j_1 + 1/2) \log B_1(j_2) + (\nu_s + 1/2) \log B_3(j_2)\right\}.$$

For the case in which $\nu_s \geq 2$, use the inequality $0 < x^{-1} \log(1 + x) < 1$ to see that $g(1)$, $g(\nu_s - 1)/(\log \nu_s)$, and $g(\nu_s)/(\log \nu_s)$ are bounded. Moreover, a calculus exercise will show that $g(j_2)$ is convex across $1 \leq j_2 \leq \nu_s - 1$. (To see this, write $g(j_2) = g_1(j_2) + g_2(j_2)$ and observe that each of $g_1(j_2)$ and $g_2(j_2)$ are convex.) It follows that $g(j_2) \leq g^* = \max\{g(1), g(\nu_s - 1)\}$ across $j_2 = 1, \ldots, \nu_s - 1$. Hence, by evaluating the "inner sum" in parts, one has

$$\begin{aligned} \sum_{j_2=0}^{\nu_s} U^{in}_{j_2}(j_1) \;\; &\leq \;\; 1 + B_0 \sum_{j_2=1}^{\nu_s - 1} \frac{1}{j_2!} \exp\left[j_2 \left\{\log \zeta_n + 2 \log p_n - 2 + g^*\right\}\right] + U^{in}_{\nu_s}(j_1) \\ &\leq \;\; 1 + B_0 \exp\left\{\log \zeta_n + 2 \log p_n - 2 + g^*\right\} + U^{in}_{\nu_s}(j_1), \end{aligned} \tag{36}$$

having noted that the middle term evaluates the first few terms of the power series expansion of the exponential function. Because $g^* = O(\log p_n)$ and $g(\nu_s) = O(\log p_n)$, as has been shown, the conditions of the theorem imply that the $\log \zeta_n$ term dominates in each exponent of the last two terms in (36), the exponent of $U^{in}_{\nu_s}(j_1)$ defined in (35), sending the exponent diverging toward negative infinity. It follows that the "inner sum" converges, $\sum_{j_2=0}^{\nu_s} U^{in}_{j_2}(j_1) \to 1$. The same conclusion holds for the case $\nu_s \leq 1$, and is deduced from (36) in a parallel way by noting that the middle term is not present when $\nu_s = 1$ and neither of the last two terms are present when $\nu_s = 0$.

Having bounded the "inner sum," it is straightforward to bound the "outer sum" using familiar techniques. It has been shown that $R^*_s \leq B_5 \sum_{j_1=0}^{k_s} U^{out}_{j_1}$, eventually, for some fixed constant B_5. Subsequently, upon noting that $U^{out}_{j_1} = (1/j_1!)\tau^{j_1}$, by simple cancellation, the "outer sum" is understood from the power series expansion of the exponential function,

$$\sum_{j_1=0}^{k_s} U^{out}_{j_1} \;\; = \;\; \sum_{j_1=0}^{k_s} \frac{1}{j_1!} \tau^{j_1} \leq e^{\tau}.$$

which establishes that R_s^* is bounded and completes the proof. □

Proof. (THEOREM 2) Write $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{Z}}(\tilde{\boldsymbol{\theta}}|\boldsymbol{\phi})$ for the respective values of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{Z}(\hat{\boldsymbol{\theta}}|\boldsymbol{\phi})$ calculated at the solution $\tilde{\boldsymbol{Y}}$.

For any sequence $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$, because $\hat{\boldsymbol{I}}_n(\boldsymbol{\theta}_n|\boldsymbol{\phi})$ is a constant multiple of a derivative of $\tilde{\boldsymbol{Z}}(\boldsymbol{\theta}_n|\boldsymbol{\phi})$, the two objects diverge at the same rate, as $n \to \infty$, unless $\tilde{\boldsymbol{Z}}(\boldsymbol{\theta}_n|\boldsymbol{\phi}) \to 0$, which requires $\boldsymbol{\theta}_n \to \boldsymbol{\theta}_0$. Apply the approximation (18) to rewrite (19) as

$$|\hat{\boldsymbol{I}}_n(\tilde{\boldsymbol{\theta}}|\boldsymbol{\phi})|^{1/2} \approx e^{\frac{1}{2}\|\boldsymbol{Z}(\tilde{\boldsymbol{\theta}}|\boldsymbol{\phi})\|^2}(2\pi)^{\nu/2}\pi(\tilde{\boldsymbol{\theta}}|\boldsymbol{\phi}) \tag{37}$$

The failure of $\tilde{\boldsymbol{Z}}(\boldsymbol{\theta}|\boldsymbol{\phi}) \to 0$ would imply the contradictory situation wherein the left and right sides of (37) are asymptotically dissimilar. It must therefore be the case that $\tilde{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0$.

By (18), the solution (19) satisfies

$$e^{\frac{1}{2}\|\tilde{\boldsymbol{Z}}(\boldsymbol{\phi})\|^2} \approx \tilde{\tau}^{\nu}\frac{|\hat{\boldsymbol{I}}_n(\tilde{\boldsymbol{\theta}}|\boldsymbol{\phi})|^{1/2}}{(2\pi)^{\nu/2}\pi(\tilde{\boldsymbol{\theta}}/\tilde{\tau}|\boldsymbol{\phi})},$$

and the approximate Bayes factor is,

$$BF_{st}(\tilde{\boldsymbol{Y}}|\boldsymbol{\phi}) \approx \left(\frac{\tau}{\tilde{\tau}}\right)^{\nu}\frac{\pi(\tilde{\boldsymbol{\theta}}/\tilde{\tau}|\boldsymbol{\phi})}{\pi(\tilde{\boldsymbol{\theta}}/\tau|\boldsymbol{\phi})} \approx \left(\frac{\tau}{\tilde{\tau}}\right)^{\nu}\frac{\pi(\boldsymbol{\theta}_0/\tilde{\tau}|\boldsymbol{\phi})}{\pi(\boldsymbol{\theta}_0/\tau|\boldsymbol{\phi})} \approx \left(\frac{\tau}{\tilde{\tau}}\right)^{\nu}$$

□

Acknowledgements

The author is grateful for invaluable support in preparing this article from the National Science Foundation (grant number SES-1260803) and the National Institute of Standards and Technology's Center for Statistics and Applications in Forensic Evidence.

References

Bartlett, M. S., (1957), Comment on D. V. Lindleys statistical paradox. *Biometrika*, 44:533-534.

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics*, 40:1550-1577.

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.

Berger, J. and Pericchi, L., (2004), Training Samples in Objective Model Selection, *Annals of Statistics*, 32:841-869.

Berry, D. A., and Hochberg, Y. (1999), Bayesian perspectives on multiple comparisons, *Journal of Statistical Planning and Inference*, 82:215-227.

Berry, S. M., and Berry, D. A. (2004), Accounting for multiplicities in assessing drug safety: a three-level hierarchical mixture model, *Biometrics*, 60:418-426.

Billingsley, P. (1995). *Probability and measure*. 3rd ed. New York: Wiley.

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.

Casella, G., Girón, F. J., Martinez, M. L., and Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *Annals of Statistics*, 37:1207-1228.

Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015), Bayesian linear regression with sparse priors, *Annals of Statistics*, 43, 1986-2018.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2012), Joint specification of model space and parameter space prior distributions, *Statistical Science*, 27:232-246.

Fan, J., and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society B*, 70:849-911.

Fan, J., and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101-148.

Fouskakis, D., Ntzoufras, I. and Perrakis, K. (2017), Power-expected-posterior priors for generalized linear models. *Bayesian Analysis*, to appear. Advanced publication at `https://projecteuclid.org/euclid.ba/1507341641`

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

Ibrahim, J. G. and Chen, M. H. (2000), Power prior distributions for regression models, *Statistical Science*, 15:46-60.

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.

Lavine, M., and Schervish, M. J., (1999) Bayes factors: what they are and what they are not, *The American Statistician,* 53:119-122.

Liang, F., Paulo, R., Molina, G., Clyde, C. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103: 410-423.

Lindley, D.V. (1957). A statistical paradox. *Biometrika*, 44: 187-192.

Lindley D.V. (1977) A problem in forensic science. *Biometrika*, 64: 207-213.

Lund, S.P., and Iyer, H., (2017). Likelihood ratio as weight of forensic evidence: a closer look. *Journal of Research of National Institute of Standards and Technology*. Preprint available at `https://arxiv.org/abs/1704.08275`

Moreno, E. and Pericchi, L. R., (2014), Intrinsic priors for objective Bayesian model selection. In *Bayesian Model Comparison (I. Jeliazkov and D. J. Poirier, eds.)*, Emerald Group Publishing Limited, 279-300.

Müller, P., Parmigiani, G., and Rice, K. (2007), FDR and Bayesian multiple comparison rules, *Bayesian Statistics 8*, (eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West), Oxford: Oxford University Press, 349-370.

Narisetty, N. N., and He, X. (2014), Bayesian variable selection with shrinking and diffusing priors. *Annals of Statistics*, 42: 789-817

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

Pérez, J. M., and Berger, J. O. (2002), Expected-posterior prior distributions for model selection, *Biometrika,* 89:491-511.

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.

Robert, C. P. (1993), A note on Jeffreys-Lindley paradox, *Statistica Sinica*, 3:603-608.

Robert, C. P., and Casella, G. (1999), *Monte Carlo Statistical Methods*, New York: Springer.

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

Scott, J. G., and Berger, J. O. (2010). Bayes and empirical Bayes multiplicity adjustment in the variable selection problem. *Annals of Statistics*, 38:2587-2619.

Spitzner, D. J. (2008), An asymptotic viewpoint on high-dimensional Bayesian testing, *Bayes-ian Analysis*, 3:121-160.

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

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

Wilson, M. A., Iversen, E. S., Clyde, M. A., Schmidler, S. C. and Schild-

kraut, J. M. (2010), Bayesian model search and multilevel inference for SNP association studies, *Annals of Applied Statistics* 4:1342-1364.

Womack, A. J., Fuentes, C. and Taylor-Rodriguez, D. (2015), Model space priors for objective sparse Bayesian regression, preprint available at `https://arxiv.org/abs/1511.04745`.

Zellner, A. (1986), On assessing prior distributions and Bayesian regression analysis using g-prior distributions, in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti (P. Goel and A. Zellner, eds)*, Amsterdam: North-Holland, 233-243