Bayesian Analysis (0000) **00**, Number 0, pp. 1

What differential geometry says about Bayesian marginalization

D.A.S. Fraser* and Mylène Bédard†‡

Abstract. The statistical tool box offers many methods for applied statistics, but reliability in the sense of 'reproducibility of frequency properties' can often be unclear or even ignored. We examine this for default Bayes methods and develop a prior that leads to full second-order inference for any regular scalar parameter of interest in presence of nuisance parameters; the new prior is Jeffreys based. Also, in parallel, we show that such second-order accuracy is widely unavailable for vector parameters of interest by Bayes, unless the interest parameter has a special linearity. Detailed examples, including simulations, are presented and discussed.

MSC 2010 subject classifications: Primary 62F15; secondary 62E20.

Keywords: coverage, exponential model, Jeffreys, mathematical prior, reproducibility.

1 Introduction

As part of discussing Lindley's (Lindley, 1975) view of a Bayesian 21st century, Efron (2013) proposed a simple classification for Bayes priors: 'genuine priors' where there is a valid frequency source for the parameter value, and 'uninformative or mathematical priors' for formal calculations without such a valid source. The conditional probability lemma of course says that when parameter values are actually sourced from a specified prior and the model itself is valid, then the posterior distribution exhibits the stated frequency property for the parameter. But in cases where other priors describe the sourcing, then the lemma says nothing in terms of frequency properties for the particular context. Our view is that such subjective priors (mathematical or other) should be recorded along side the confidence-based presentation, so that an end user knows the sources of the information presented and can reweight posterior draws to explore alternative analyses. This is not to say that such posteriors might not have attractive properties; for example, if a quantile bound obtained from some posterior exhibits the frequency-under-repetitions claimed by its labeling, then that frequency is de facto confidence for the quantile bound.

Berger (2006) and Goldstein (2011) refer to the unification of Bayesian and frequentist procedures in terms of coverage-matching. This property arises when a Bayes calculation is examined under repetitions and found to exhibit the stated posterior

*Department of Statistical Sciences, University of Toronto, Ontario, Canada, M5S 3G3. dfraser@utstat.toronto.edu

†Département de mathématiques et de statistique, Université de Montréal, Québec, Canada, H3C 3J7. bedard@dms.umontreal.ca

‡Corresponding author

 DOI: 0000

value; such a concept of course is just confidence with an alternative labelling. This indicates that a Bayes calculation can provide a route to confidence but needs checking. This viewpoint is also largely in conformity with Laplace (1812) who didn't have the 'confidence' concept available, but did have major experience in science and the remarkable insight of a great scientist. Over the years, various researchers have made progress on achieving reproducibility through Bayes. Tibshirani (1989), for instance, proposed a coverage-matching prior for the case where a scalar parameter of interest is orthogonal to the complementing nuisance parameter. More recently, Fraser et al. (2010) developed a prior based on matching Bayesian and frequentist higher-order approximations; we refer the reader to that article and the references therein for more details on the subject. Datta and Sweeting (2005) also provide a comprehensive review of available coverage-matching priors.

In this paper, we develop a new prior that leads to second-order accuracy in terms of frequency reproducibility for a scalar parameter of interest. The construction of this new prior, which was foreseen in the explorations of Fraser et al. (2016b), builds on regular models and likelihood asymptotics. Its development requires useful properties stemming from exponential models in their canonical form and, as such, involves several reparameterizations and changes of variables. This however does not jeopardize the applicability of the new prior, which can be obtained from general regular models, not only exponential ones. We record, in §2, the distributional results that are necessary to the development of the new prior.

Then in §3 we introduce the prior, which is in fact just Jeffreys prior (Jeffreys, 1946) but curiously used 'off-label', strictly on the one-dimensional profile contour for the interest parameter. In the case where the interest parameter is non-linear with respect to the canonical parameter, then a Jacobian allowance is needed and uses a rotationally symmetric reparameterization of the model. We present a clear and systematic approach for computing the prior, waving the need for potentially restrictive nuisance correction terms that are contingent on the geometry of the interest parameter with respect to the canonical parameterization, see Fraser et al. (2016b). The resulting single-dimensional posterior, which may be seen as emerging from the new Jeffreys-based prior combined to the profile likelihood of the interest parameter, can then be used for second-order Bayesian inference. Following this line of reasoning, further Bayesian calculations are then accessible through the use of the one-dimensional statistical model that is proportional to the profile likelihood of interest (instead of the initial full-dimensional statistical model).

In parallel we also show, through a revealing example in §4, that posteriors for vector parameters quite generally do not have such confidence accuracy, particularly when marginalized to component parameters. In fact, priors featuring second-order accuracy are widely unavailable for vector parameters of interest, unless the parameter has a special linearity. Finally, in §5, we examine a spectrum of examples in details, and find that the new prior gives remarkable accuracy for posterior quantile bounds and intervals.

2 Some results on regular models and likelihood asymptotics

Consider a statistical model $f(y;\theta)$ with observed data y^0, where y has dimension n and θ has dimension p. In recent years, many of the most productive developments for statistical analysis have come from the saddlepoint approximation as promoted by Daniels (1954). This method offers a highly accurate $\mathcal{O}(n^{-3/2})$ reexpression of an exponential family model, hereafter referred to as third-order accurate. Let s be a canonical variable and φ be a canonical parameter for such a model in exponential form: $f(y;\theta) = \exp[\varphi(\theta)^\top s(y) - \kappa\{\varphi(\theta)\}]h(y)$, where $\varphi(\theta) \in \mathbb{R}^p$ is one-to-one-equivalent to θ and $s(y) \in \mathbb{R}^p$. The approximate model can then be presented entirely in terms of very familiar statistical quantities as

$$\begin{aligned} g(s;\varphi)\mathrm{d}s &= \exp\{\varphi^\top s - \kappa(\varphi)\}g(s)\mathrm{d}s \\ &= \frac{\mathrm{e}^{k/n}}{(2\pi)^{p/2}} \exp\{\ell(\varphi;s) - \ell(\hat{\varphi};s)\}|\hat{\jmath}_{\varphi\varphi}|^{-1/2}\{1+\mathcal{O}(n^{-3/2})\}\mathrm{d}s \ , \end{aligned} \tag{1}$$

where $\ell(\varphi;s) - \ell(\hat{\varphi};s) = -r^2/2$ is the negative log-likelihood ratio, $\hat{\varphi} = \hat{\varphi}(s)$ is the maximum likelihood estimator, and $\hat{\jmath}_{\varphi\varphi} = \jmath_{\varphi\varphi}(\hat{\varphi})$ is the observed information matrix in the canonical parameterization. Each of these involves dependence on the variable s but only the first has dependence also on φ. The term k/n is a generic normalizing constant. The high accuracy of the approximate model is important, but pales in contrast to the ability to extract definitive departure measures of data from parameter, essentially replacing any use of sufficiency, ancillarity and other reduction methods, yet retaining continuity in wide generality. Hereafter, we refer to the approximation in (1) (minus the error term) as $g(s;\varphi)$; this sort of gives the final null distribution $g(s;\varphi_0)$ for assessing $\varphi = \varphi_0$ in a single, unequivocal step. We note that for more general regular models (not in exponential form), there exists an effective construct for such a φ so the present methodology becomes widely available; see Appendix A.

Now to begin suppose we have a scalar-variable, scalar-parameter model in exponential form. An intriguing result from Welch and Peers (1963) shows that the approximation $g(s;\varphi)$ can be rewritten as a location model, say $\bar{g}(t-\mu)$ with $t = t(s)$ and $\mu = \mu(\varphi)$, to second-order accuracy $\mathcal{O}(n^{-1})$. This is achieved by making use of Taylor expansions, information functions, and transformations on the variable and parameter. Using this convenient approximate location property, the authors then show that the observed p-value function is equal, to second-order accuracy, to the Bayes survivor function under Jeffreys (1946) prior. In particular,

$$\int^{s^0} \frac{\mathrm{e}^{k/n}}{(2\pi)^{1/2}} \mathrm{e}^{-r_{\psi_0}^2/2} \jmath_{\varphi\varphi}^{-1/2}(\hat{\varphi})\mathrm{d}s = \int_{\varphi_0} \frac{\mathrm{e}^{k/n}}{(2\pi)^{1/2}} \mathrm{e}^{-r_\psi^2/2} \jmath_{\varphi\varphi}^{1/2}(\varphi)\mathrm{d}\varphi \ , \tag{2}$$

where the left term is the observed p-value function $p(\varphi_0) = \int^{s^0} g(s;\varphi_0)\mathrm{d}s$ and the expression on the right is the Bayes survivor function $s(\varphi_0) = \int_{\varphi_0} \pi(\varphi|s^0)\mathrm{d}\varphi$ as based on Jeffreys' prior $\pi(\varphi) \propto \jmath_{\varphi\varphi}^{1/2}(\varphi)$. Both sides of (2) are expressed in terms of $r_\psi^2/2$, the constrained signed log-likelihood ratio $\ell(\hat{\varphi}; s) - \ell(\hat{\varphi}_\psi; s)$, where $\hat{\varphi}_\psi$ is the constrained maximum likelihood estimator given the interest parameterization $\psi(\varphi) = \psi$. On the left, the interest ψ_0 is fixed and $r^2_{\psi_0}/2 = \ell(\hat{\varphi}; s) - \ell(\hat{\varphi}_{\psi_0}; s)$, whereas on the right, the data is at its observed value and $r^2_\psi/2 = \ell(\hat{\varphi}; s^0) - \ell(\hat{\varphi}_\psi; s^0)$. The p-value essentially records the statistical position of the data relative to a parameter value φ_0 such that $\psi(\varphi_0) = \psi_0$; alternatively, it could be interpreted as the smallest significance level to which one would reject the null in testing $H_0 : \psi(\varphi) = \psi_0$ against $H_1 : \psi(\varphi) < \psi_0$. In the current context, we can say that the root information prior gives Bayes-frequency equivalence.

More generally, suppose we have a p-dimensional exponential model and are interested in a scalar parameter $\psi = \psi(\varphi)$; let $\lambda = \lambda(\varphi)$ be a $(p-1)$-dimensional complementing nuisance parameter for ψ. It is convenient to assume that the nuisance λ is chosen orthogonal to the scalar interest ψ in the sense of Cox and Reid (1987). This constraint will however be swallowed up in the theoretical developments of §3 and will not be required in the examples or, more generally, in practice. As it turns out, the unique null distribution for assessing a particular ψ value is directly available from asymptotic theory; we now provide the broad lines of the argument, see Fraser and Reid (1995) or Fraser (2016) for more details. With $\psi(\varphi)$ fixed at ψ_0, there exists an approximate ancillary statistic U for the nuisance parameter, i.e. a function of s whose distribution is second-order free of λ. This statistic $U(s)$ takes values on a continuous contour in the sample space; $U(s)$ and the contour on which it is defined may not be unique, but the ancillary distribution is unique to the third order. For convenience, let the continuous contour be the observed profile line $L^0_{\psi_0} = \{s \in \mathbb{R}^p : \hat{\lambda}_{\psi_0} = \hat{\lambda}^0\}$, on which the constrained maximum likelihood estimator of λ is fixed at its observed value.

In this setting, we can reparameterize the exponential model such that $(u(s), v(s))$ acts as the full canonical variable. If the interest ψ is a linear function of φ, then an appropriate change of variable easily leads to an exponential model with canoncial parameter $(\psi(\varphi), \lambda(\varphi))$ and canonical variable $(u(s), v(s))$; from there, a saddlepoint approximation $\tilde{g}(u, v; \psi, \lambda)$ as in (1) is then accessible. If the interest is not linear in φ, then we define a new parameter $\chi = \chi(\varphi)$ linear in φ and tangent to $\psi(\varphi)$ at $\hat{\varphi}_{\psi_0}$, the constrained full maximum likelihood value given $\psi(\varphi) = \psi_0$. Through a change of variable, we then obtain an exponential model with canonical variable $(u(s), v(s))$ and canonical parameter $\chi(\varphi)$ instead of $\psi(\varphi)$ (in addition to the complementing nuisance parameter). We note that on $L^0_{\psi_0}$, the original exponential model coincides with this tangent exponential model, for which a saddlepoint approximation is also available; the analysis would then be pursued with the latter, but for simplicity we hold on to our usual notation ψ for the interest parameter.

In particular, let (u, v) be the full canonical variable and consider an approximating exponential model as in (1); keeping $\psi(\varphi)$ fixed at ψ_0, we can reexpress the full model as

$$\tilde{g}(u, v; \psi_0, \lambda) = q(v|u; \psi_0, \lambda) h(u; \psi_0) \ ,$$

with a nuisance density $q(v|u; \lambda, \psi_0)$ and an interest density $h(u; \psi_0)$ that contains full third-order information on ψ_0. To obtain an expression for $h(u; \psi_0)$ on $L^0_{\psi_0}$, we need to isolate it in the previous equation. The statistic U being ancillary, the observed profile line $L^0_{\psi_0}$ corresponds to the line on which the $(p-1)$-dimensional variable v is fixed at its observed value v^0. On $L^0_{\psi_0}$, we thus have access to an approximating exponential model $\tilde{g}(u, v^0; \psi_0, \hat{\lambda}^0)$ for the full density; in φ parameterization, this gives

$$\begin{aligned}\tilde{g}(u, v^0; \hat{\varphi}_{\psi_0})\mathrm{d}u &= \frac{\mathrm{e}^{k/n}}{(2\pi)^{p/2}} \exp\{\ell(\hat{\varphi}_{\psi_0}; u, v^0) - \ell(\hat{\varphi}; u, v^0)\}|\jmath_{\varphi\varphi}\{\hat{\varphi}(u, v^0)\}|^{-1/2}\mathrm{d}u \\ &= \frac{\mathrm{e}^{k/n}}{(2\pi)^{p/2}} \mathrm{e}^{-r^2_{\psi_0}/2}|\jmath_{\varphi\varphi}\{\hat{\varphi}\}|^{-1/2}\mathrm{d}u \ .\end{aligned}$$

The conditional density of v given the ancillary $U = u$ and its contour L_{ψ_0} is $q(v|u; \psi_0, \lambda)$, which inherits exponential form and thus also admits the saddlepoint approximation; evaluated at v^0 and expressed in terms of φ, it satisfies

$$q(v^0|u; \hat{\varphi}_{\psi_0}) = \frac{\mathrm{e}^{k/n}}{(2\pi)^{(p-1)/2}} \exp\{0\}|\jmath_{(\lambda\lambda)}(\hat{\varphi}_{\psi_0})|^{-1/2} \ .$$

The parentheses around $\lambda\lambda$ indicate that the second derivative must be rescaled to that of the given exponential parameterization φ, thus making the expression free of the nuisance λ. The nuisance information determinant in the parameterization (λ) can be obtained from the determinant in the parameterization λ by applying the Jacobian $\varphi_\lambda = \partial\varphi/\partial\lambda$ for fixed ψ,

$$|\jmath_{(\lambda\lambda)}(\hat{\varphi}_\psi)| = |\jmath_{\lambda\lambda}(\hat{\varphi}_\psi)||\varphi_\lambda^\top(\hat{\varphi}_\psi)\varphi_\lambda(\hat{\varphi}_\psi)|^{-1} \ .$$

Dividing the full density by the conditional one leads to the marginal null distribution on $L^0_{\psi_0}$, with parameter ψ_0 and scalar differential $\mathrm{d}u$

$$h(u; \psi_0)\mathrm{d}u = \frac{\mathrm{e}^{k/n}}{(2\pi)^{1/2}} \mathrm{e}^{-r^2_{\psi_0}/2}|\jmath_{\varphi\varphi}(\hat{\varphi})|^{-1/2}|\jmath_{(\lambda\lambda)}(\hat{\varphi}_{\psi_0})|^{1/2}\mathrm{d}u \ . \tag{3}$$

The above is similar to (1), except that it also involves the observed nuisance information determinant $|\jmath_{(\lambda\lambda)}(\hat{\varphi}_{\psi_0})|$. The expression (3) also happens to be valid for vector ψ. We remind the reader that in all previous expressions, k is a generic normalizing constant. For more about the development of $h(u; \psi)$, and in particular about the implications of using the tangent exponential model, see Fraser (2011).

From the orthogonality between ψ and λ, the preceding null distribution can be rearranged using a factorization of the determinant $|\jmath_{\varphi\varphi}| = |\jmath_{(\lambda\lambda)}||\jmath^{(\psi\psi)}|^{-1}$, where a full matrix with upper indices is the inverse of the same with lower indices. The marginal null distribution in (3) becomes

$$h(u; \psi_0)\mathrm{d}u = \frac{\mathrm{e}^{k/n}}{(2\pi)^{1/2}} \mathrm{e}^{-r^2_{\psi_0}/2} \cdot \left\{ \frac{|\jmath_{(\lambda\lambda)}(\hat{\varphi}_{\psi_0})|}{|\jmath_{(\lambda\lambda)}(\hat{\varphi})|} \right\}^{1/2} \cdot |\jmath^{(\psi\psi)}(\hat{\varphi})|^{1/2}\mathrm{d}u \ . \tag{4}$$

Suppose we temporarily ignore the factor in $\{\}$; the contribution $|\jmath^{(\psi\psi)}(\hat{\varphi})|^{1/2}\mathrm{d}u$ then appears as the (Welch and Peers, 1963) differential on the sample space line $L^0_{\psi_0}$

with respect to an underlying scalar exponential model. This presents the contribution $e^{k/n}(2\pi)^{-1/2}\exp\{-r^2_{\psi_0}/2\}|\jmath^{(\psi\psi)}(\hat\varphi)|^{1/2}\mathrm{d}u$ as a location model to second-order accuracy, as in (2). Now the factor $\{|\jmath_{(\lambda\lambda)}(\hat\varphi_{\psi_0})|/|\jmath_{(\lambda\lambda)}(\hat\varphi)|\}^{1/2}$, already second-order accurate, can be expanded as a function $\exp\{a(t-\mu)/n^{1/2}\}$, with t,μ as previously mentioned (Fraser et al., 2016b). The combination then is a function of $(t-\mu)$, providing a full location form for (4) to second order. The result of Welch and Peers (1963) can thus be applied on this distribution, but some technicalities need attention beforehand.

The distribution (3) is on the line L^0_ψ for a fixed ψ and goes through the observed data; it is also perpendicular to the interest parameter contour at the constrained maximum likelihood value $\hat\varphi_\psi$ on the parameter space. The parameter $\psi(\varphi)$ often has certain rotation properties; accordingly, the line L^0_ψ could change direction with ψ-change. As a result, the observed information on the line L^0_ψ could vary and, correspondingly, so could the form of the underlying apparent exponential distribution. We can notationally avoid this complication in the use of Welch and Peers (1963) by recalibrating the exponential coordinates to have an observed information matrix equal to the identity, $\hat\jmath_{\varphi\varphi}=I$. This is not a change in substance, just notational so that what we have written as an exponential model is, under rotation, the same exponential model to second order. The recalibration is achieved by using a right square root T of $\hat\jmath_{\varphi\varphi}=T^\top T$ and then using the modified canonical parameter $\bar\varphi=T\varphi$ that has now acquired an identity observed information. This means that the reference exponential distributions through the data point are now, notationally, a single exponential distribution.

3 A Jeffreys-based prior featuring second-order reproducibility

We now combine the previous distributional results with Welch and Peers (1963)'s relationship to derive a prior that achieves confidence. Hereafter, we use the modified exponential parameterization $\bar\varphi$ as just described at the end of §2; for notational simplicity, we however write φ and assume that the adjustment has been made. The density (4) can be integrated up to the observed $u=u^0$, leading to the distribution function $H(u^0;\psi_0)$ called the p-value function $p(\psi_0)$:

$$p(\psi_0)=\int^{u^0}\frac{e^{k/n}}{(2\pi)^{1/2}}e^{-r^2_{\psi_0}/2}\cdot\left\{\frac{|\jmath_{(\lambda\lambda)}(\hat\varphi_{\psi_0})|}{|\jmath_{(\lambda\lambda)}(\hat\varphi)|}\right\}^{1/2}\cdot|\jmath^{(\psi\psi)}(\hat\varphi)|^{1/2}\mathrm{d}u\ ,$$

where $r^2_{\psi_0}$ and the information functions of course depend on the scalar u. Recall the uniqueness of this distribution subject to retaining model continuity and containing full information for the parameter ψ_0.

Then applying Welch and Peers (1963)'s location model result at (2) to the location model obtained in (4), we can rewrite the p-value function (an integral on the observed profile line $L^0_{\psi_0}$ with fixed parameter ψ_0) as a survivor posterior function (an integral

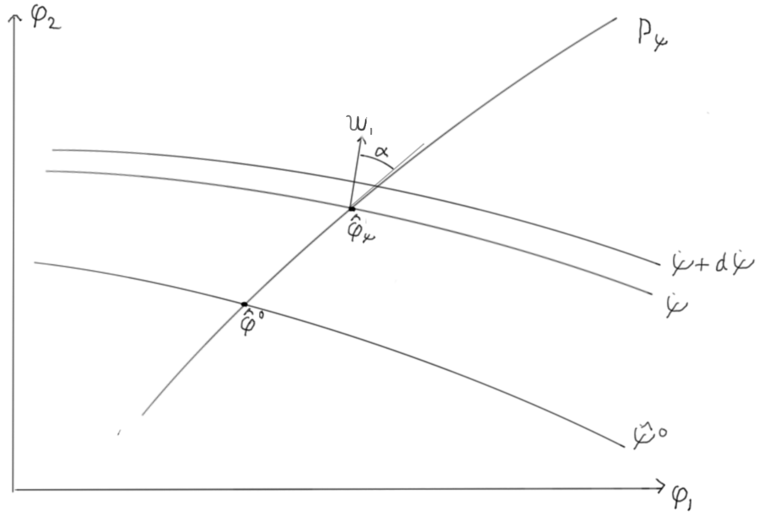

Figure 1: Exponential coordinates having symmetry ($\hat{\jmath}^0_{\varphi\varphi} = I$) at the point $\hat{\varphi}^0$; $\mathrm{d}\psi$ is an increment in the parameter ψ; $\mathrm{d}\hat{\varphi}_\psi$ is the corresponding vector increment for the point $\hat{\varphi}_\psi$ on the profile curve P_ψ; and $\mathrm{d}(\psi)$ is the corresponding increment in the symmetrized exponential coordinates.

on the profile curve $P_\psi = \{\psi \in \mathbb{R} : \varphi = \hat{\varphi}_\psi\}$, with $\hat{\varphi}_\psi = \hat{\varphi}^0_\psi$ based on observed data s^0):

$$p(\psi_0) \quad = \quad \int_{\psi_0} \frac{\mathrm{e}^{k/n}}{(2\pi)^{1/2}} \mathrm{e}^{-r_\psi^2/2} \cdot \quad \left\{ \frac{|\jmath_{(\lambda\lambda)}(\hat{\varphi}_\psi)|}{|\jmath_{(\lambda\lambda)}(\hat{\varphi})|} \right\}^{1/2} \cdot |\jmath^{(\psi\psi)}(\hat{\varphi}_\psi)|^{-1/2} \ \mathrm{d}(\psi) \ .$$

We remind the reader that the term $\{|\jmath_{(\lambda\lambda)}(\hat{\varphi}_\psi)|/|\jmath_{(\lambda\lambda)}(\hat{\varphi})|\}^{1/2}$ has location form to second-order accuracy, which is required for the application of that result. We can then combine the information functions that depend on $\hat{\varphi}_\psi$ into a single factor, and absorb the information that depends only on the data into the constant k:

$$p(\psi_0) = \int_{\psi_0} \frac{\mathrm{e}^{k/n}}{(2\pi)^{1/2}} \mathrm{e}^{-r_\psi^2/2} \cdot |\jmath_{\varphi\varphi}(\hat{\varphi}_\psi)|^{1/2} \ \mathrm{d}(\psi) \ . \tag{5}$$

As before, the parentheses around ψ indicate that the parameterization is in its exponential version on the profile curve; more details about the differential $\mathrm{d}(\psi)$ are provided below. Furthermore, to avoid notational difficulties with parameter rotation, we have required the exponential parameter φ to be locally rotationally symmetric as described in the last paragraph of §2.

The $\mathcal{O}(n^{-1})$ version of the p-value function $p(\psi_0)$ in (5) has now been written as an integral of observed likelihood on the parameter space. But, this integral of likelihood is totally restricted to the profile curve P_ψ; as such, the integral is a contour integral, and not the usual full parameter space integral! The integrand in (5) can thus be seen as a posterior density for ψ obtained from Jeffreys' prior used 'off-label', on just the profile P_ψ for the interest parameter. Furthermore, this modification with focus on the interest parameter has full second-order repetition accuracy.

We now study the support differential $\mathrm{d}(\psi)$ based on the special exponential parameterization (ψ) for the interest ψ. Consider how a change $\mathrm{d}\psi$ affects the constrained full maximum likelihood value $\hat{\varphi}_\psi$ along the profile; this is given by the derivative $\mathrm{d}\hat{\varphi}_\psi/\mathrm{d}\psi$, which has magnitude $|\mathrm{d}\hat{\varphi}_\psi/\mathrm{d}\psi|$. Then, let α be the angle between the profile contour P_ψ and the gradient vector $\mathrm{d}\psi/\mathrm{d}\varphi$ of the ψ surface; see Figure 1. The change perpendicular to a ψ contour is then obtained by multiplying the preceding magnitude by the cosine of that angle

$$\begin{aligned} \mathrm{d}(\psi) &= |\frac{\mathrm{d}\hat{\varphi}_\psi}{\mathrm{d}\psi}| \ \cos\{\alpha\} \ \mathrm{d}\psi \\ &= |\frac{\mathrm{d}\hat{\varphi}_\psi}{\mathrm{d}\psi}| \ w_1 w_2 \ \mathrm{d}\psi \\ &= w_1 \ \frac{\mathrm{d}\hat{\varphi}_\psi}{\mathrm{d}\psi} \ \mathrm{d}\psi \ , \end{aligned} \tag{6}$$

where $w_1 = \{\mathrm{d}\psi(\varphi)/\mathrm{d}\varphi\}/|\mathrm{d}\psi(\varphi)/\mathrm{d}\varphi|$ is the unit gradient vector to the surface $\psi(\varphi) = \psi$ at $\hat{\varphi}_\psi$ and $w_2 = \{\mathrm{d}\hat{\varphi}_\psi/\mathrm{d}\psi\}/|\mathrm{d}\hat{\varphi}_\psi/\mathrm{d}\psi|$ is the unit vector associated to the above-mentioned change in $\hat{\varphi}_\psi$ along the profile.

We thus obtain, to second order, the Bayes posterior survivor value function for a general scalar interest parameter ψ_0:

$$p(\psi_0) = c \ \int_{\psi_0} \mathrm{e}^{-r_\psi^2/2} |\jmath_{\varphi\varphi}(\hat{\varphi}_\psi)|^{1/2} \cdot |\frac{\mathrm{d}\hat{\varphi}_\psi}{\mathrm{d}\psi}| \ \cos\{\alpha\} \ \mathrm{d}\psi \ ,$$

where $|\mathrm{d}\hat{\varphi}_\psi/\mathrm{d}\psi| \cos\{\alpha\}$ represents the Jacobian for Jeffreys' prior on the profile curve, $|\jmath_{\varphi\varphi}(\hat{\varphi}_\psi)|^{1/2}$. The implicit prior density is thus expressed as

$$\pi_D(\psi) \ \mathrm{d}\psi \propto |\jmath_{\varphi\varphi}(\hat{\varphi}_\psi)|^{1/2} \cdot w_1 \frac{\mathrm{d}\hat{\varphi}_\psi}{\mathrm{d}\psi} \cdot \mathrm{d}\psi \ , \tag{7}$$

for ψ on P_ψ and may be used with the profile log-likelihood $\ell(\hat{\varphi}_\psi; s^0)$ for further Bayesian developments. Note that the nuisance parameter nowhere appears in the prior nor the posterior; in practice, it is thus not required to identify a nuisance λ that is orthogonal to ψ. The posterior survivor function in this section has second-order uniqueness and accuracy by its derivation from the p-value function, which in turn has second-order uniqueness and accuracy by calculation respecting continuity. For some related discussion, see Fraser (2014).

4 A prior for sets?

In §2 and §3 we considered vector parameters, from which we then identified a scalar parameter of interest. By focussing on this interest parameter we found that a prior with second-order reproducibility is available, on just the profile contour for that parameter. A reasonable concern is whether such second-order posterior accuracy might be available more generally, for say a compact set. For this, we work with a simple core case where we can reasonably hope that things would be easy: let $S = (S_1, S_2)$ on the plane be standard Normal with distribution located at $\varphi = (\varphi_1, \varphi_2)$; this is a simple exponential model with canonical variable s and canonical parameter φ. Despite the simple location form of this model, we will see that Welch and Peers (1963)'s result does not hold for sets and therefore Bayes survivor functions do not generally have reproducibility.

First let $\rho^2 = \varphi_1^2 + \varphi_2^2$ be the parameter of interest; the set $\{s \in \mathbb{R}^2 : s_1^2 + s_2^2 < r^2\}$, where say $r^2 = (s_1^0)^2 + (s_2^0)^2$ is the observed statistics, is then standard for calculating the p-value function. The p-value $p(\rho_0^2) = \mathbb{P}(S_1^2 + S_2^2 < r^2; \rho_0^2)$ associated to a specific ρ_0^2 is then given by the Non-central Chi-squared distribution function $H_2(r^2; \rho_0^2)$. The usual Bayes survivor function, obtained with Jeffreys' flat prior for location parameters, is then given by $1 - H_2(\rho_0^2; r^2)$. Hence, sample space probability within the disk with radius r (using the distribution with parameter ρ_0^2) can be compared with the Bayes parameter space probability outside the circle with radius ρ_0 (using the distribution with parameter r^2). It follows that the Bayes posterior survivor function is larger than the p-value function, a familiar result in the presence of parameter curvature. And then if we decrease the value of r, the p-value moves towards 0 and the survivor function moves towards 1. In the extreme, the p-value can be close to 0 and the corresponding survivor value can be close to 1. As the p-value has repetition validity it follows that the Bayes survivor probability in general does not, here to the extreme.

Figure 2 illustrates the behaviour of the p-value and posterior survivor value functions of the parameter ρ^2 for various values of the radius ($r^2 = 4, 2, 1, 0.5$). In particular, the discrepancy between both approaches becomes larger as r^2 decreases, in which case the p-value function becomes closer to 0. The above should not be surprising given the behaviour of the pivotal r/ρ in calculating confidence.

5 Examples

We now present a range of examples based on simple exponential models in which the parameter of interest ψ increases in complexity. We begin with a parameter ψ that is linear in the canonical parameterization φ, then study a rotational ψ, followed by a curved one, and we finally address the Behrens-Fisher problem. We detail how the new reproducibility prior may be obtained in each of these cases, and graphically assess its performance by comparison to the frequentist benchmark that is the third-order p-value.

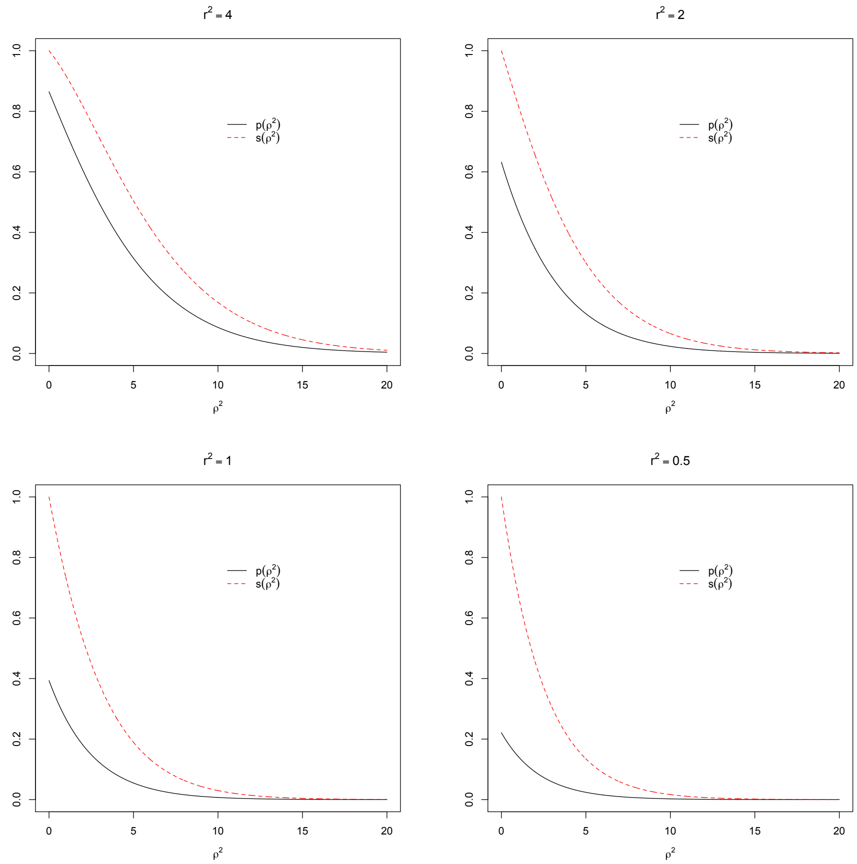

Figure 2: P-value and posterior survivor value functions for the parameter ρ^2; each graph has its own r^2 value: 4, 2, 1, and 0.5.

5.1 Linear Parameter

Consider an interest parameter ψ linear in the canonical parameterization φ, i.e. $\psi(\varphi) = v^\top\varphi = \sum v_i\varphi_i$. This is the simplest case as the line L^0_ψ then remains parallel to the vector v under ψ changes. Since L^0_ψ does not rotate, there is no need for invoking rotational symmetry in the observed information matrix $\hat{\jmath}_{\varphi\varphi}$, thus waiving the recalibration discussed at the end of §2. Users looking for an automated implementation of the method could however include a default use of this recalibration without altering the results.

Specifically, consider a beta density with canonical parameter $\varphi = (\alpha, \beta)$:

$$f(y;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} \ , \quad y \in (0,1) \ ,$$

with $n = 5$ observed values $\mathbf{y}^0 = (0.36, 0.68, 0.44, 0.43, 0.34)$. The parameter α is of interest ($\psi = \alpha$), while β is a free nuisance ($\lambda = \beta$). The interest ψ is linear in the canonical parameterization, as $\psi(\varphi) = v^\top \varphi = \alpha$ with $v^\top = (1, 0)$.

We aim at comparing p-value functions $p(\alpha)$ to posterior survivor value functions $s(\alpha)$ arising from available uninformative priors. To this end, the signed log-likelihood root approach is used as a simple approximation to $p(\alpha)$, while the third-order approach acts as a highly accurate approximation (Fraser, 2017). These are then compared to posterior survivor value functions $s(\alpha)$ obtained using Jeffreys' prior and the new directional Jeffreys-style prior.

The beta model does not admit a closed-form expression for its maximum likelihood estimates (MLEs). Using the function `beta.mle` in the `R` package `Rfast` leads to $\hat{\varphi}^0 = (7.47, 9.03)$; these are used in approximating the p-value function $p(\alpha)$. The constrained MLE of β given α, $\hat{\beta}_\alpha$, is the solution of

$$D'(\hat{\beta}_\alpha) - D'(\alpha + \hat{\beta}_\alpha) = \frac{1}{n} \sum_{i=1}^{n} \log(1 - y_i) \ ,$$

where $D'(x) = \mathrm{d} \log \Gamma(x)/\mathrm{d}x$ is the digamma function. This equation is solved using the function `uniroot` in `R`; constrained MLEs $\hat{\beta}_\alpha$ are obtained for various values of the interest α, and then used in approximating $p(\alpha)$ and computing posterior survivor values $s(\alpha)$ based on the new directional Jeffreys-style prior.

The Fisher information function appears in every calculation except that of the signed log-likelihood root; it satisfies

$$\jmath_{\varphi\varphi}(\varphi) \quad = \quad \begin{pmatrix} n\,(D''(\alpha) - D''(\alpha+\beta)) & -nD''(\alpha+\beta) \\ -nD''(\alpha+\beta) & n\,(D''(\beta) - D''(\alpha+\beta)) \end{pmatrix} ,$$

where $D''(x) = \mathrm{d}^2 \log \Gamma(x)/\mathrm{d}x^2$ is the trigamma function. Jeffreys' prior does not distinguish between interest and nuisance parameters; it is defined on the full parameter space as the root of the Fisher information determinant:

$$\pi_J(\alpha, \beta) \ \propto \ \{D''(\alpha)D''(\beta) - D''(\alpha+\beta)\,[D''(\alpha) + D''(\beta)]\}^{1/2} \ , \quad \alpha, \beta > 0 \ .$$

We note that the Bayesian benchmark prior, the reference prior of Bernardo (1979), is not easily available for a beta model in which an interest parameter is targeted. If it were, it would also lead to a prior on the full parameter space, but in which interest and nuisance parameters have been treated differently in the derivation of the density.

As a new way of targeting the interest parameter, the directional Jeffreys-style prior restricts the usual Jeffreys' prior to the profile contour for the interest α. From (7), the new prior π_D satisfies

$$\pi_D(\alpha)\ \mathrm{d}\alpha \quad \propto \quad |\jmath_{\varphi\varphi}(\hat{\varphi}_\psi)|^{1/2}\ \mathrm{d}(\alpha) \ = \ \pi_J(\alpha, \hat{\beta}_\alpha)\ \mathrm{d}(\alpha) \ .$$

In the current simple linear case, w_1 in (6) is the vector $(1, 0)$, and thus

$$\mathrm{d}(\alpha) \ = \ w_1 \, \frac{\mathrm{d}(\alpha, \hat{\beta}_\alpha)^\top}{\mathrm{d}\alpha} \, \mathrm{d}\alpha \ = \ \mathrm{d}\alpha \ .$$

The posterior survivor value function $s_D(\alpha)$ is then obtained by integrating the one-dimensional posterior density

$$\pi_D(\alpha|\mathbf{y}^0)\ \mathrm{d}\alpha \quad \propto \quad \exp\left\{\ell(\alpha,\hat{\beta}_\alpha;\mathbf{y}^0)\right\}|\jmath_{\varphi\varphi}(\alpha,\hat{\beta}_\alpha)|^{1/2}\ \mathrm{d}\alpha\ ,$$

where $\ell(\alpha,\hat{\beta}_\alpha;\mathbf{y}^0) = \log f(\mathbf{y}^0;\alpha,\hat{\beta}_\alpha)$ denotes the profile log-likelihood function of the interest α.

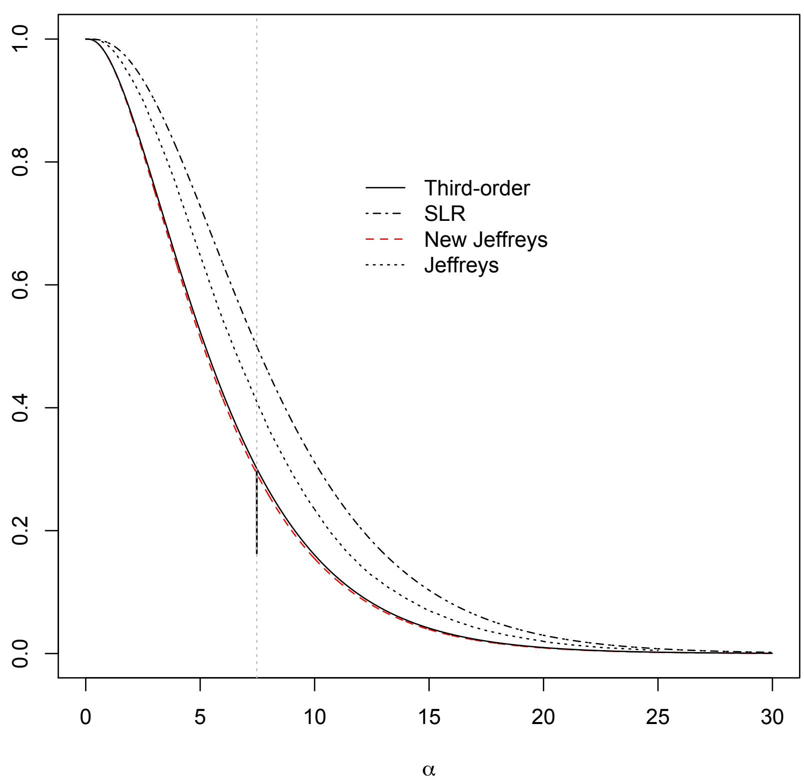

Figure 3: P-value and posterior survivor value functions for the parameter α in the beta model; the MLE of α is identified by a pale vertical line.

Figure 3 examines the third-order function $p(\alpha)$ taken as the exact p-value (solid line) and the normal approximation for the signed log-likelihood root r_α (dash-dotted line). The graph also features a comparison with posterior survivor values obtained under Jeffreys' prior (dotted line) and the new directional Jeffreys (red dashed line). Approximations of the p-value function have been obtained in `R`, while the posterior survivor values were obtained by running 100,000 iterations of a random walk Metropolis algorithm with a Gaussian proposal distribution featuring a scaling $\sigma^2 = 4$ (also in `R`). In the current context, the directional Jeffreys offers second-order reproducibility; this is not available from the regular Jeffreys, which treats both parameters as of equal importance.

In this example, we studied the simplistic case where $\varphi = (\alpha, \beta) = \theta = (\psi, \lambda)$. Linear examples with $\varphi \neq \theta$ are easy to find and in such cases, the development of the new prior is similar to that expounded in the current section. Consider, for instance, the same beta model and let $\psi = \alpha + \beta$, $\lambda = \beta$; this yields an interest parameter that is still linear in φ, expressed as $\psi(\varphi) = v^\top \varphi$ with $v^\top = (1, 1)$. This time, the constrained MLE of β given ψ, $\hat{\beta}_\psi$, is the solution of

$$D'(\psi - \hat{\beta}_\psi) - D'(\hat{\beta}_\psi) \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \log y_i - \frac{1}{n} \sum_{i=1}^{n} \log(1 - y_i) \ .$$

The vector w_1 in (6) is $w_1 = (1, 1)/\sqrt{2}$, leading to

$$\mathrm{d}(\psi) = w_1 \, \frac{\mathrm{d}(\psi - \hat{\beta}_\psi, \hat{\beta}_\psi)^\top}{\mathrm{d}\psi} \, \mathrm{d}\psi \ . \tag{8}$$

In practice, an analytical expression for $\mathrm{d}\hat{\varphi}_\psi/\mathrm{d}\psi$ is not always available. In such cases, $\mathrm{d}(\psi)$ is simply reexpressed as $\mathrm{d}(\psi) = w_1 \mathrm{d}\hat{\varphi}_\psi$ and posterior survivor values are then easily obtained from numerical integration, by selecting an appropriately small lag h and letting $\mathrm{d}\hat{\varphi}_\psi \approx \hat{\varphi}_{\psi+h} - \hat{\varphi}_\psi$.

From (7) and (8), the new prior satisfies $\pi_D(\psi)\mathrm{d}\psi \propto |j_{\varphi\varphi}(\psi - \hat{\beta}_\psi, \hat{\beta}_\psi)|^{1/2} w_1 \mathrm{d}\hat{\varphi}_\psi$ and is used along with the profile likelihood to obtain posterior survivor values, as explained above. Figure 4 provides a comparison similar to that found in Figure 3, outlining virtually parallel performances amongst implemented methods.

5.2 Rotating Parameter

In several cases, the direction of the line L^0_ψ may vary under ψ changes. Although this does not happen in linear cases, more generally, L^0_ψ may rotate through an $\mathcal{O}(n^{-1/2})$ angle. This even happens in very simple settings and with classical distributions, as the following example illustrates.

Consider a normal model in which $Y \sim \mathcal{N}(\mu, \sigma^2)$ and let $\theta = (\psi, \lambda) = (\mu, \sigma^2)$. For a vector of n observations, the log-likelihood function of this model satisfies

$$\ell(\mu, \sigma^2; \mathbf{y}) \quad = \quad -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} y_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} y_i - \frac{n\mu^2}{2\sigma^2} \ . \tag{9}$$

From (9), the canonical parameters are $\varphi(\theta) = (\mu/\sigma^2, -1/\sigma^2)$. The interest parameter thus satisfies $\psi(\varphi) = -\varphi_1/\varphi_2 = \mu$, which is obviously not linear in φ. The maximum

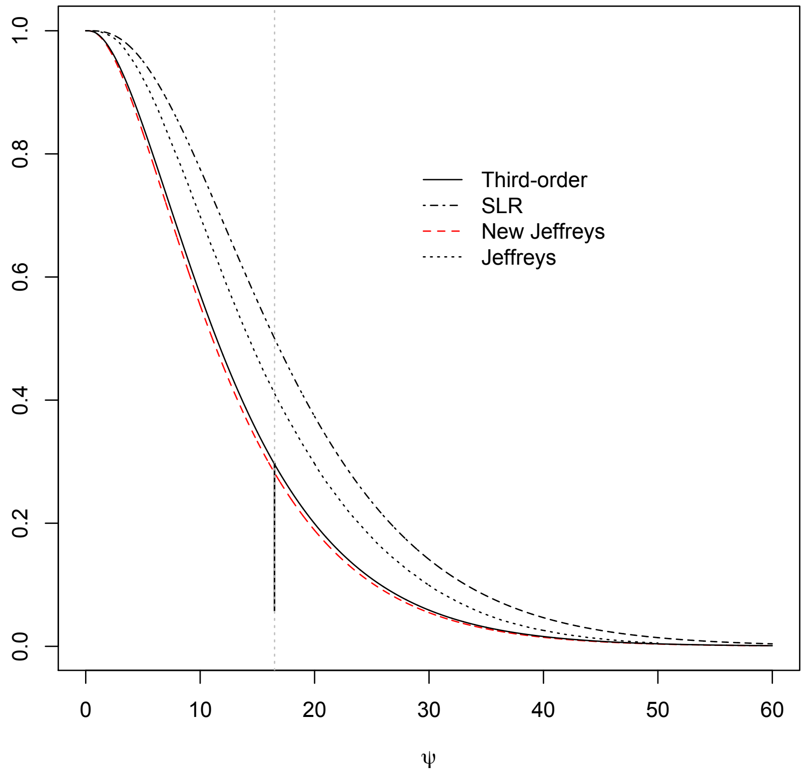

Figure 4: P-value and posterior survivor value functions for the parameter $\psi = \alpha + \beta$ in the beta model; the MLE of ψ is identified by a pale vertical line.

likelihood estimates in the canonical parameterization are $\hat{\varphi} = (n\bar{y}/S^2, -n/S^2)$, where $\bar{y} = \sum_{i=1}^n y_i/n$ and $S^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2$. The constrained MLE for σ^2 given μ is $\hat{\sigma}^2_\mu = \{S^2 + n(\bar{y}-\mu)^2\}/n$, leading to $\hat{\varphi}_\psi = (\mu/\hat{\sigma}^2_\mu, -1/\hat{\sigma}^2_\mu)$. The Fisher information function is expressed as

$$\jmath_{\varphi\varphi}(\varphi) = \begin{pmatrix} -n/\varphi_2 & n\varphi_1/\varphi_2^2 \\ n\varphi_1/\varphi_2^2 & n/2\varphi_2^2 - n\varphi_1^2/\varphi_2^3 \end{pmatrix} . \tag{10}$$

Using (10), Jeffreys' prior is $\pi_J(\varphi)\mathrm{d}\varphi \propto |\jmath_{\varphi\varphi}(\varphi)|^{1/2}\mathrm{d}\varphi \propto (-\varphi_2)^{-3/2}\,\mathrm{d}\varphi$ on $\mathbb{R}\times\mathbb{R}^-$. The reference prior satisfies $\pi_R(\varphi)\mathrm{d}\varphi \propto \mathrm{d}\varphi/\varphi_2^2$, see Bernardo (1979).

We now proceed to determine the new Jeffreys-style prior, based on an observed sample $\mathbf{y}^0 = (0.00, 1.10, -0.50, 0.25, -0.95, -0.60,\ 0.35)$. Since the angle of L^0_ψ rotates under ψ changes, we apply a recalibration $\bar{\varphi} = T\varphi$ with $\jmath_{\varphi\varphi}(\hat{\varphi}^0) = T^\top T$, as mentioned in §2. For simplicity and interchangeability in the use of T and its transpose, we find the eigenvalues and eigenvectors of $\jmath_{\varphi\varphi}(\hat{\varphi}^0)$ with the function `eigen` in `R`, and then use these quantities to define a symmetrical matrix T.

In practice, this change from φ to $\bar{\varphi}$ only affects the differential $\mathrm{d}(\psi)$ in (7). Indeed, since $\jmath_{\bar{\varphi}\bar{\varphi}} = (T^{-1})^\top \jmath_{\varphi\varphi} T^{-1}$, then $|\jmath_{\bar{\varphi}\bar{\varphi}}| = |\jmath_{\varphi\varphi}|/|T|^2$; when evaluated at $\theta = (\psi, \hat{\lambda}_\psi)$ along the profile curve, these determinants are proportional with respect to ψ and therefore interchangeable in terms of Bayesian computations. Now, the determinant of (10) evaluated at $\hat{\varphi}_\psi$ that appears in 7 is computed as

$$|\jmath_{\varphi\varphi}(\hat{\varphi}_\psi)|^{1/2} \quad \propto \quad \left(\hat{\sigma}^2_\mu\right)^{3/2} \;\propto\; \{S^2 + n(\bar{y}-\mu)^2\}^{3/2} \; . \tag{11}$$

We finally develop the differential term $\mathrm{d}(\psi)$ in (6). It is crucial to explicitly take account of the recalibration T in this Jacobian. The term $\mathrm{d}\psi/\mathrm{d}\bar{\varphi}$ in (6) satisfies

$$\frac{\mathrm{d}\psi}{\mathrm{d}\bar{\varphi}} \quad = \quad \frac{\mathrm{d}\psi(\varphi)}{\mathrm{d}\varphi}\,\frac{\mathrm{d}\varphi}{\mathrm{d}\bar{\varphi}} \;=\; \left(-\frac{1}{\varphi_2}, \frac{\varphi_1}{\varphi_2^2}\right) T^{-1} \;=\; \sigma^2(1,\mu)T^{-1} \; ,$$

which can then be normalized to the unit vector $w_1 = \{\mathrm{d}\psi/\mathrm{d}\bar{\varphi}\}/|\mathrm{d}\psi/\mathrm{d}\bar{\varphi}|$ and evaluated at $\hat{\varphi}_\psi$. The term $\mathrm{d}(T\hat{\varphi}_\psi)/\mathrm{d}\psi$ in (6) is obtained as

$$\begin{aligned} T\frac{\mathrm{d}\hat{\varphi}_\psi}{\mathrm{d}\psi} \quad &= \quad T\frac{\mathrm{d}}{\mathrm{d}\mu}\left(\frac{\mu}{\hat{\sigma}^2_\mu}, -\frac{1}{\hat{\sigma}^2_\mu}\right)^\top \\ &= \quad T\frac{1}{(\hat{\sigma}^2_\mu)^2}\left(\hat{\sigma}^2_\mu + 2\mu(\bar{y}-\mu), -2(\bar{y}-\mu)\right)^\top \; . \end{aligned}$$

This finally leads to

$$\begin{aligned} \mathrm{d}(\psi) \quad &= \quad \left|\frac{\mathrm{d}\psi}{\mathrm{d}\bar{\varphi}}\right|^{-1} \frac{\mathrm{d}\psi}{\mathrm{d}\varphi}T^{-1}\; T\frac{\mathrm{d}\hat{\varphi}_\psi}{\mathrm{d}\psi}\;\mathrm{d}\psi \\ &= \quad \frac{1}{|(1,\mu)T^{-1}|(\hat{\sigma}^2_\mu)^2}\;(1,\mu)\left(\hat{\sigma}^2_\mu + 2\mu(\bar{y}-\mu), -2(\bar{y}-\mu)\right)^\top \mathrm{d}\mu \\ &= \quad \frac{1}{|(1,\mu)T^{-1}|\hat{\sigma}^2_\mu}\;\mathrm{d}\mu \; , \end{aligned}$$

and the matrix T conveniently appears in the vector norm only. Using the latter along with (11), we obtain the directional Jeffreys-style prior

$$\pi_D(\mu)\mathrm{d}\mu \quad \propto \quad \{S^2 + n(\bar{y}-\mu)^2\}^{3/2}\;\mathrm{d}(\mu) \;\propto\; \frac{\{S^2 + n(\bar{y}-\mu)^2\}^{1/2}}{|(1,\mu)T^{-1}|}\;\mathrm{d}\mu \; ;$$

the resulting posterior survivor value is

$$s_D(\mu_0) \quad = \quad \int_{\mu_0} \exp\left\{\ell(\mu, \hat{\sigma}^2_\mu; \mathbf{y}^0)\right\}\;\pi_D(\mu)\;\mathrm{d}\mu \; .$$

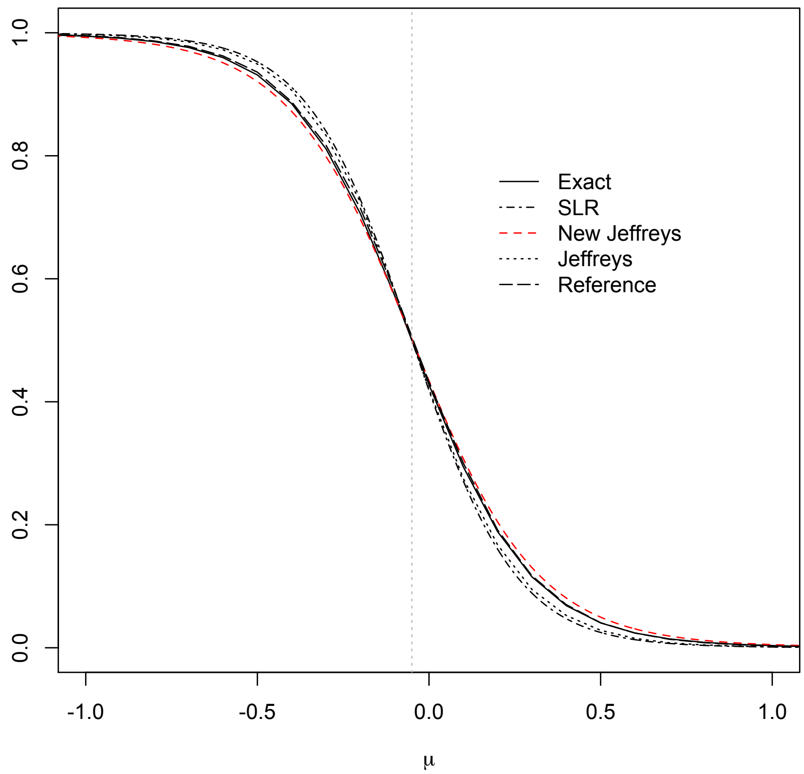

Figure 5: P-value and posterior survivor value functions for the parameter μ in the normal model; the MLE of μ is identified by a pale vertical line.

Figure 5 examines p-value and posterior survivor value functions obtained with the observed sample $\mathbf{y}^0$. The exact p-value function $p(\alpha)$ is obtained using a Student-t distribution with $n-1$ degrees of freedom and is represented on the graph by a solid line. The normal approximation for the signed likelihood root is also included (dash-dotted line). The graph features a comparison with posterior survivor values obtained under Jeffreys' prior (dotted line), the reference prior (long-dash), and the new directional Jeffreys (red dashed line). Exact and approximated p-value functions have been obtained in `R`, while the posterior survivor values (based on Jeffreys and reference) were obtained by running 200,000 iterations of a random walk Metropolis algorithm with a Gaussian proposal distribution featuring a scaling $\sigma^2 = 0.40$ (also in `R`). Posterior survivor values using the new directional Jeffreys were obtained through numerical integration. Results from the new Jeffreys-style prior are as convincing as those based on the Bayesian benchmark, the reference prior.

5.3 Curved Parameter

As an example with curvature, consider a gamma model with canonical parameters $\alpha, \beta > 0$. We are interested in the variance $\psi = \alpha/\beta^2$, which is curved in terms of $\varphi = (\alpha, \beta)$, and we choose to work with the free nuisance parameter $\lambda = \beta$. The density of the model is

$$f(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp\{-\beta y\} \ , \quad y > 0 \ ,$$

with $n = 5$ observed values $\mathbf{y}^0 = (0.20, 0.45, 0.78, 1.28, 2.28)$ as used in Brazzale et al. (2007) on page 13. The maximum likelihood estimates of the canonical parameters, $\hat{\varphi} = (\hat{\alpha}, \hat{\beta})$, are the solution of the equations

$$\begin{aligned} \hat{\alpha}/\hat{\beta} &= \bar{y} \ , \\ D'(\hat{\alpha}) - \log \hat{\alpha} &= \frac{1}{n} \sum_{i=1}^{n} \log y_i - \log \bar{y} \ . \end{aligned}$$

By re-expressing the log-likelihood function in terms of interest and nuisance as $\ell(\psi, \lambda; \mathbf{y})$, we find the constrained MLE $\hat{\lambda}_\psi$ to be the solution of

$$2\psi\lambda \left[\log \lambda + \frac{1}{2} - D'(\psi\lambda^2) + \frac{1}{n} \sum_{i=1}^{n} \log y_i \right] = \bar{y} \ .$$

The Fisher information function in the canonical parameterization is

$$\jmath_{\varphi\varphi}(\varphi) = \begin{pmatrix} nD''(\alpha) & -n/\beta \\ -n/\beta & n\alpha/\beta^2 \end{pmatrix} \ ,$$

and so Jeffreys' prior $|\jmath_{\varphi\varphi}(\varphi)|^{1/2}$, which treats both parameters as of equal interest, is $\pi_J(\varphi)\mathrm{d}\varphi \propto \{\alpha D''(\alpha) - 1\}^{1/2}/\beta \mathrm{d}\varphi$. The reference prior for this specific context would target the interest parameter $\psi = \alpha/\beta^2$, but is not widely available is this case.

Since the model studied does not satisfy the linearity constraint, a recalibration $\bar{\varphi} = T\varphi$ of the canonical parameter is required, where T is such that $\jmath_{\varphi\varphi}(\hat{\varphi}^0) = T'T$. From §5.2, this recalibration impacts the value of the differential $\mathrm{d}(\psi)$, but only through the term $|\mathrm{d}\psi/\mathrm{d}\bar{\varphi}|$. Jeffreys' prior evaluated on the profile, i.e. at $\hat{\varphi}_\psi = (\hat{\alpha}_\psi, \hat{\beta}_\psi) = (\psi\hat{\beta}_\psi^2, \hat{\beta}_\psi)$, is

$$|\jmath_{\bar{\varphi}\bar{\varphi}}(\hat{\bar{\varphi}}_\psi)|^{1/2} \ \propto \ |\jmath_{\varphi\varphi}(\hat{\varphi}_\psi)|^{1/2} \quad \propto \quad \{\hat{\alpha}_\psi D''(\hat{\alpha}_\psi) - 1\}^{1/2}/\hat{\beta}_\psi \ .$$

The term $\mathrm{d}\psi/\mathrm{d}\bar{\varphi}$ in (6) is

$$\frac{\mathrm{d}\psi}{\mathrm{d}\bar{\varphi}} \quad = \quad \left(-\frac{1}{\beta^2}, -\frac{2\alpha}{\beta^3} \right) T^{-1} \ ;$$

evaluated at $\hat{\varphi}_\psi$, it becomes $(-1/\hat{\beta}_\psi^2, -2\psi/\hat{\beta}_\psi)\ T^{-1}$. Since we did not obtain a closed-form expression for $\hat{\varphi}_\psi$, the differential term $\mathrm{d}\hat{\varphi}_\psi/\mathrm{d}\psi$ in (6) cannot be computed explicitly. In that case, we simply use the differential $\mathrm{d}(\psi) = |\mathrm{d}\psi/\mathrm{d}\bar{\varphi}|^{-1}\ \mathrm{d}\psi/\mathrm{d}\varphi\ \mathrm{d}\hat{\varphi}_\psi$, and numerically evaluate this expression for an appropriately small lag h, by letting $\mathrm{d}\hat{\varphi}_\psi \approx \hat{\varphi}_{\psi+h} - \hat{\varphi}_\psi$. This leads to the directional Jeffreys-style prior satisfying

$$\begin{aligned}\pi_D(\psi)\ \mathrm{d}\psi &\propto \pi_J(\hat{\varphi}_\psi)\ \mathrm{d}(\psi)\\ &\propto \frac{1}{\hat{\beta}_\psi}\{\hat{\alpha}_\psi D''(\hat{\alpha}_\psi)-1\}^{1/2}\cdot\frac{1}{|\mathrm{d}\psi/\mathrm{d}\varphi\ T^{-1}|}\frac{\mathrm{d}\psi}{\mathrm{d}\varphi}\ \mathrm{d}\hat{\varphi}_\psi\\ &\propto \frac{1}{\hat{\beta}_\psi}\{\hat{\alpha}_\psi D''(\hat{\alpha}_\psi)-1\}^{1/2}\cdot\frac{1}{|(1,2\psi\hat{\beta}_\psi)T^{-1}|}(1,2\psi\hat{\beta}_\psi)\ \mathrm{d}(\hat{\alpha}_\psi,\hat{\beta}_\psi)^\top ,\end{aligned}$$

and to a posterior survivor function

$$s_D(\psi_0) = \int_{\psi_0} \exp\left\{\ell(\hat{\alpha}_\psi,\hat{\beta}_\psi;\mathbf{y})\right\}\ \pi_D(\psi)\ \mathrm{d}\psi .$$

Figure 7 compares approximations of the p-value function (SLR, third-order) and posterior survivor value functions under different priors (regular Jeffreys and new directional Jeffreys). The new directional prior is again extremely close to the third-order p-value function, while Jeffreys' prior now significantly underestimates the latter.

5.4 Behrens-Fisher problem

Consider two independent variables, $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, and observe $\mathbf{y} = (\mathbf{y}_1^0, \mathbf{y}_2^0)$, with $\mathbf{y}_i^0$ of size n_i from Y_i. The interest parameter is $\psi = \mu_1 - \mu_2$ and we let the nuisance be $\lambda = (\mu_2, \sigma_1^2, \sigma_2^2)$; the full parameter is then $\theta = (\psi, \lambda)$. The log-likelihood function satisfies

$$\begin{aligned}\ell(\theta;\mathbf{y}) = &-\frac{n_1}{2}\log(2\pi\sigma_1^2) - \frac{n_2}{2}\log(2\pi\sigma_2^2)\\ &-\frac{1}{2\sigma_1^2}\{n_1(\bar{y}_1-\psi-\mu_2)^2+S_1^2\} - \frac{1}{2\sigma_2^2}\{n_2(\bar{y}_2-\mu_2)^2+S_2^2\} ,\end{aligned}$$

with $\bar{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$ and $S_i^2 = \sum_{j=1}^{n_i} y_{ij}^2 - n_i(\bar{y}_i)^2$. This leads to the full MLE $\hat{\theta} = (\bar{y}_1 - \bar{y}_2, \bar{y}_2, \frac{S_1^2}{n_1}, \frac{S_2^2}{n_2})$. To obtain the constrained MLE of λ given ψ, we solve the following system of equations:

$$\begin{aligned}\hat{\mu}_2 =& \frac{n_1\hat{\sigma}_2^2(\bar{y}_1-\psi)+n_2\hat{\sigma}_1^2\bar{y}_2}{n_1\hat{\sigma}_2^2+n_2\hat{\sigma}_1^2} \qquad (12)\\ \hat{\sigma}_i^2 =& (\bar{y}_i - \psi\cdot\mathbb{1}_{(i=1)} - \hat{\mu}_2)^2 + \frac{S_i^2}{n_i} , \quad i=1,2 ,\end{aligned}$$

where $\mathbb{1}_{(\cdot)}$ is the indicator function; plugging $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ into $\hat{\mu}_2$, we numerically solve for $\hat{\mu}_2$ and then work backwards for the variances.

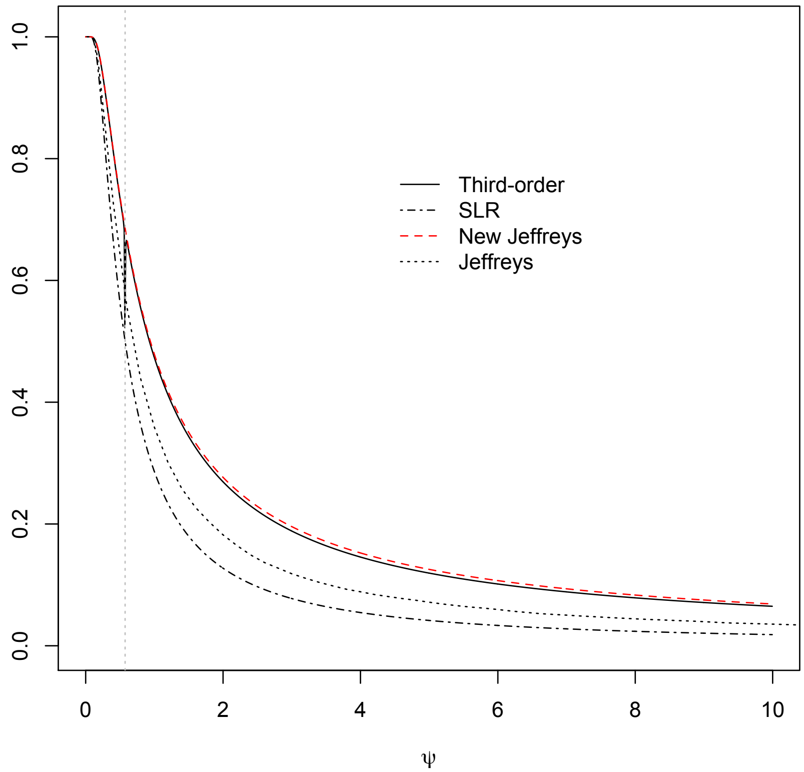

Figure 6: P-value and posterior survivor value functions for the parameter $\psi = \alpha/\beta^2$ in the gamma model; the MLE of ψ is identified by a pale vertical line.

The canonical parameter of this model is $\varphi(\theta) = (\frac{\psi+\mu_2}{\sigma_1^2}, \frac{\mu_2}{\sigma_2^2}, \frac{1}{\sigma_1^2}, \frac{1}{\sigma_2^2})$. The MLE is $\hat{\varphi} = \varphi(\hat{\theta}) = (\frac{n_1\bar{y}_1}{S_1^2}, \frac{n_2\bar{y}_2}{S_2^2}, \frac{n_1}{S_1^2}, \frac{n_2}{S_2^2})$ and the constrained MLE given ψ is $\hat{\varphi}_\psi = (\frac{\psi+\hat{\mu}_2}{\hat{\sigma}_1^2}, \frac{\hat{\mu}_2}{\hat{\sigma}_2^2}, \frac{1}{\hat{\sigma}_1^2}, \frac{1}{\hat{\sigma}_2^2})$, using estimates in (12). The log-likelihood function can be reexpressed as $\ell(\varphi; \mathbf{y})$, which leads to the information matrix

$$\jmath_{\varphi\varphi}(\varphi) = \begin{pmatrix} \frac{n_1}{\varphi_3} & 0 & -\frac{n_1\varphi_1}{\varphi_3^2} & 0 \\ 0 & \frac{n_2}{\varphi_4} & 0 & -\frac{n_2\varphi_2}{\varphi_4^2} \\ -\frac{n_1\varphi_1}{\varphi_3^2} & 0 & \frac{n_1}{2\varphi_3^2} + \frac{n_1\varphi_1^2}{\varphi_3^3} & 0 \\ 0 & -\frac{n_2\varphi_2}{\varphi_4^2} & 0 & \frac{n_2}{2\varphi_4^2} + \frac{n_2\varphi_2^2}{\varphi_4^3} \end{pmatrix}$$

with determinant $|\jmath_{\varphi\varphi}(\varphi)| = n_1^2 n_2^2/\{4\varphi_3^3\varphi_4^3\}$. Jeffreys' prior for this problem is therefore $\pi_J(\varphi)\mathrm{d}\varphi \propto |\jmath_{\varphi\varphi}(\varphi)|^{1/2}\mathrm{d}\varphi \propto (\varphi_3\varphi_4)^{-3/2}\mathrm{d}\varphi$, while the reference prior satisfies

$\pi_R(\varphi)\mathrm{d}\varphi \propto (\varphi_3\varphi_4)^{-2}\mathrm{d}\varphi$.

We now work on finding the new prior. The interest parameter ψ is not a linear function of φ, as $\psi(\varphi) = \varphi_1/\varphi_3 - \varphi_2/\varphi_4$. We therefore need to recalibrate and work with $\bar{\varphi} = T\varphi$, where $\jmath_{\varphi\varphi}(\hat{\varphi}) = T^\top T$. This transformation only has an impact on the differential $d(\psi)$ and does not affect the term $\pi_J(\hat{\varphi}_\psi) = |\jmath_{\varphi\varphi}(\hat{\varphi}_\psi)|^{1/2}$. The differential $\mathrm{d}\psi/\mathrm{d}\bar{\varphi}$ is

$$\frac{\mathrm{d}\psi}{\mathrm{d}\bar{\varphi}} = \left(\frac{1}{\varphi_3}, -\frac{1}{\varphi_4}, -\frac{\varphi_1}{\varphi_3^2}, \frac{\varphi_2}{\varphi_4^2}\right) T^{-1} \ ,$$

which can then be normalized to the unit vector $w_1 = \{\mathrm{d}\psi/\mathrm{d}\bar{\varphi}\}/|\mathrm{d}\psi/\mathrm{d}\bar{\varphi}|$ and evaluated at $\hat{\varphi}_\psi$. Since we did not obtain a closed-form expression for $\hat{\varphi}_\psi$, the differential term $\mathrm{d}T\hat{\varphi}_\psi/\mathrm{d}\psi$ in (6) cannot be computed explicitly. In that case, we simply use the differential $\mathrm{d}(\psi) = w_1\ T\ \mathrm{d}\hat{\varphi}_\psi$, and numerically evaluate this expression for an appropriately small lag h, by letting $\mathrm{d}\hat{\varphi}_\psi \approx \hat{\varphi}_{\psi+h} - \hat{\varphi}_\psi$. This leads to the directional Jeffreys-style prior satisfying

$$\begin{aligned} \pi_D(\psi)\ \mathrm{d}\psi &\propto \pi_J(\hat{\varphi}_\psi)\ \mathrm{d}(\psi) \\ &= (\hat{\sigma}_1^2\hat{\sigma}_2^2)^{3/2} \cdot \frac{1}{|\mathrm{d}\psi/\mathrm{d}\varphi\ T^{-1}|}\frac{\mathrm{d}\psi}{\mathrm{d}\varphi}\ \mathrm{d}\hat{\varphi}_\psi \ , \end{aligned}$$

and to a posterior survivor function

$$s_D(\psi_0) = \int_{\psi_0} \exp\left\{\ell(\hat{\varphi}_\psi; \mathbf{y})\right\}\ \pi_D(\psi)\ \mathrm{d}\psi \ .$$

Figure 7 provides a comparison of p-value and posterior survivor value functions similar to previous examples; it is based on the dataset $\mathbf{y}_1^0 =$ (1.02, 0.82, -0.37, 0.40, 1.29, 1.39, -0.21), $\mathbf{y}_2^0 = (-0.86$, -2.13, -0.76, 0.60, 0.26, -0.74, 0.49). For the Behrens-Fisher problem, it is well-known that Jeffreys' prior leads to a p-value function that is reproducible to second-order. Naturally, as the reference prior differs from the latter, it now either over- or under-estimates the p-value, depending of the specific ψ tested. As expected, the new directional Jeffreys-based prior is extremely close to the third-order p-value, which illustrates its robustness across various contexts.

6 Discussion

Efron (2013) offered a classification of Bayes priors, mentioning 'genuine priors' when there is an objective random source for the actual parameter value, and 'uninformative priors' for formal calculations, sometimes referred to as mathematical priors. For the non-genuine priors, Berger (2006) and Goldstein (2011) recommend unifying Bayesian and frequentist procedures, by which they mean reproducibility, repetition under identical conditions. Repetition reliability has had extensive discussion in the frequency literature and leads to third-order accuracy for scalar parameters with most regular models. With this as a benchmark under repetitions, we have developed a second-order

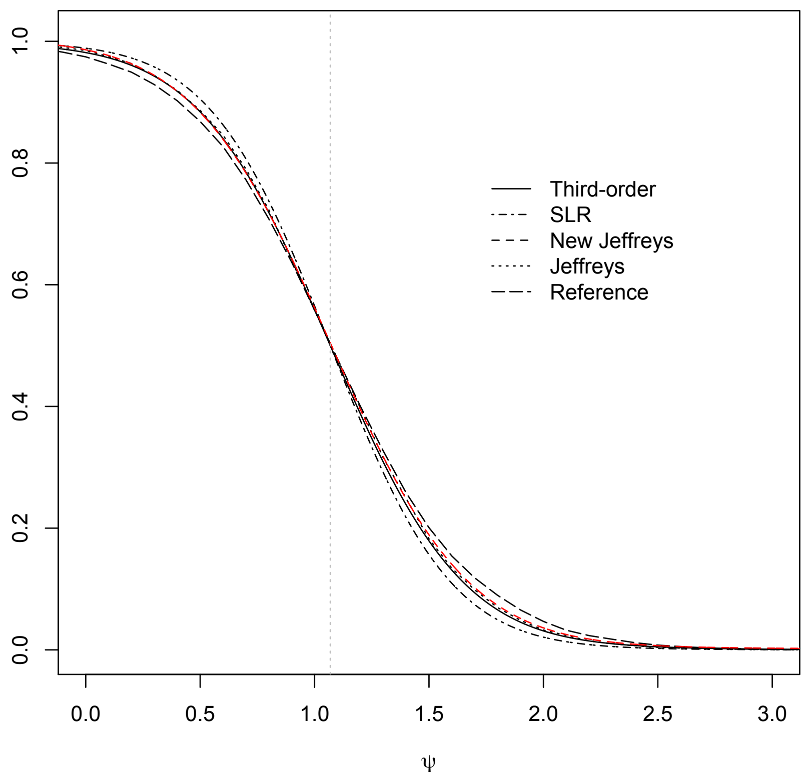

Figure 7: P-value and posterior survivor value functions for the parameter $\psi = \mu_1 - \mu_2$ in the Behrens-Fisher problem; the MLE of ψ is identified by a pale vertical line.

accurate prior for scalar parameters and find it to be essentially Jeffreys' prior but confined to the profile contour for the scalar parameter of interest; this indicates that the ordinary use of Jeffreys does what might be viewed as a double overlapping calculation.

According to the theory exposed, Bayesian inference should then use the profile likelihood with parameter ψ, along with the new Jeffreys-based prior. Although the resulting one-dimensional posterior appears to rely on plug-in estimators, we note that it arises from usual Bayesian arguments such as marginalization. Indeed an ancillary statistic u, whose distribution is free of the nuisance parameter, was first identified; an expression for the density of this statistic was then obtained by marginalizing the joint density with respect to λ. Therefore, the issue of the theoretical developments not be mistaken with a deliberate plug-in approach.

Examples have been investigated under increasing complexity: linear, rotating, curved, and they clearly support the claimed second-order accuracy. Vector interest parameters, however, do not generally have repetition reliability; this was investigated by Dawid et al. (1973) as marginalization paradoxes, and by the present discussion under parameter curvature. Second-order frequency based p-values for vector parameters are available from Fraser et al. (2016a).

A Appendix: from exponential to general models

The results discussed in this paper were presented for regular exponential models, but they are available for quite general regular models. For this consider an n-dimensional variable with a p-dimensional full parameter, plus continuity for parameter effects. In the simple scalar variable and parameter case, the distribution function $F(y;\theta) = z$ (say) can be inverted to give the quantile function $y = y(z;\theta)$. This allows easy simulations for the variable y using an underlying uniform distribution for z. The same is widely available for the vector variable case, by determining an $n \times p$ matrix

$$V = (v_1, \dots, v_p) = \frac{\partial y}{\partial \theta} \ ,$$

where the differentiation is for fixed pivotal $z = z(y;\theta) = z(y^0;\hat{\theta}^0)$. Differentiating the log-model $\ell(\theta; y)$ in the directions V then gives the needed canonical parameter

$$\varphi = \frac{\partial \ell(\theta; y)}{\partial V}\Big|_{y^0} \ ,$$

which is used with an observed canonical variable $s = 0$. This leads to an exponential model using (φ, s), called the tangent exponential model, which then provides full third-order inference for the original model data combination.

References

Berger, J. (2006). "The case for objective Bayesian analysis." *Bayesian Analysis*, 1: 385–402. 1, 20

Bernardo, J. M. (1979). "Reference posterior distributions for Bayesian inference." *J Royal Statist Soc B*, 41: 113–147. 11, 14

Brazzale, A. R., Davison, A. C., and Reid, N. (2007). *Applied Asymptotics Case Studies in Small-Sample Statistics*. Cambridge, UK: Cambridge University Press. 17

Cox, D. R. and Reid, N. (1987). "Parameter orthgonality and approximate conditional inference (with Discussion)." *J Royal Statist Soc B*, 49: 1–39. 4

Daniels, H. E. (1954). "Saddlepoint approximations in statistics." *Annals Math. Statist.*, 46: 21–31. 3

Datta, G. S. and Sweeting, T. J. (2005). "Probability matching priors." *Handbook of Statistics*, 25: 91–114. 2

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). "Marginalization paradoxes in Bayesian and structural inference." *J. Roy. Statist. Soc. B*, 35: 189–233. 22

Efron, B. (2013). "Bayes' Theorem in the 21st Century." *Science*, 340: 1177–1178. 1, 20

Fraser, D. (2011). "Assessing a value for an interest parameter and a definitive reference distribution." *Biometrika*. 5

Fraser, D., Reid, N., and Sartori, N. (2016a). "Accurate directional inference for vector parameters." *Biometrika*, 103(3): 625–639. 22

Fraser, D. A. S. (2014). "Why does statistics have two theories?" In Lin, X., Genest, C., Banks, D. L., Molenberghs, G., Scott, D. W., and Wang, J.-L. (eds.), *Past, Present and Future of Statistical Science*, 237–252. Florida: CRC Press. 8

— (2016). "Definitive testing of an interest parameter: Using parameter continuity." *Jour. Statist. Res.*, 47: 153–165. 4

— (2017). "The p-value function: The core concept of modern statistical inference." *Annual reviews in statistics and its applications*, 4: 153–165. 11

Fraser, D. A. S., Bédard, M., Wong, A., Lin, W., and Fraser, A. (2016b). "Bayes, reproducibility and the quest for truth." *Statistical Science*, 31: 578–590. 2, 6

Fraser, D. A. S. and Reid, N. (1995). "Ancillaries and third order significance." *Utilitas Mathematica*, 47: 33 – 53. 4

Fraser, D. A. S., Reid, N., Marras, E., and Yi, G. (2010). "Default priors for Bayesian and frequentist inference." *J. Royal Statist. Soc., Ser B*, 75: 631–654. 2

Goldstein, M. (2011). "Subjective Bayesian analysis: Principles and practice." *Bayesian Analysis*, 26: 187–202. 1, 20

Jeffreys, H. (1946). "An invariant form for the prior probability in estimation problems." *Proc. Roy. Soc. A*, 186: 453–461. 2, 3

Laplace, P. S. d. (1812). *Théorie Analytique des Probabilités*. Paris: Courcier. 2

Lindley, D. (1975). "The Future of Statistics: A Bayesian 21st Century." *Advances in Applied Probability*, 7: 106–115. 1

Tibshirani, R. (1989). "Noninformative priors for one parameter of many." *Biometrika*, 76(3): 604–608. 2

Welch, B. L. and Peers, H. W. (1963). "On formulae for confidence points based on intervals of weighted likelihoods." *Jour. Royal Statist. Soc. B*, 25: 318–329. 3, 5, 6, 9

Acknowledgments

We acknowledge support from the Natural Sciences and Engineering Research Council of Canada, and the Senior Scholars Funding of York University.