

The Whetstone and the Alum Block: Balanced Objective Bayesian Comparison of Nested Models for Discrete Data*

Guido CONSONNI, Jonathan J. FORSTER and Luca LA ROCCA

Abstract. When two nested models are compared, using a Bayes factor, from an objective standpoint, two seemingly conflicting issues emerge at the time of choosing parameter priors under the two models. On the one hand, for moderate sample sizes, the evidence in favor of the smaller model can be inflated by diffuseness of the prior under the larger model. On the other hand, asymptotically, the evidence in favor of the smaller model typically accumulates at a slower rate. With reference to finitely discrete data models, we show that these two issues can be dealt with jointly, by combining intrinsic priors and non-local priors in a new unified class of priors. We illustrate our ideas in a running Bernoulli example, then we apply them to test the equality of two proportions, and finally we deal with the more general case of logistic regression models.

AMS 2000 subject classifications: Primary 62F15; Secondary 62F03.

Key words and phrases: Bayes factor, Intrinsic prior, Model choice, Moment prior, Non-local prior, Ockham's razor, Training sample size.

1. INTRODUCTION

Consider two parametric models, \mathcal{M}_0 (the *null* model) nested in \mathcal{M}_1 (the *alternative* model), each equipped with its own prior distribution, $p_0(\cdot)$ and $p_1(\cdot)$. We plan to compare models using the Bayes Factor (BF); see Kass and Raftery (1995) for a classic review. We denote by $f_i(\cdot|\theta_i)$ the sampling density of data y

Guido Consonni is Professor of Statistics, Università Cattolica del Sacro Cuore, Dipartimento di Scienze Statistiche, Milano, Italy (e-mail:

guido.consonni@unicatt.it). Jonathan J. Forster is Professor of Statistics, University of Southampton, School of Mathematics and Southampton Statistical Sciences Research Institute, Southampton, UK (e-mail: j.j.forster@soton.ac.uk).

Luca La Rocca is Assistant Professor of Statistics, Università di Modena e Reggio Emilia, Dipartimento di Scienze Fisiche, Informatiche e Matematiche, Modena, Italy (e-mail: luca.larocca@unimore.it).

*This work was partially supported by MIUR, Rome, PRIN 2007XECZ7L_001, and the University of Pavia; it was started while Consonni and La Rocca were visiting the Southampton Statistical Sciences Research Institute at the University of Southampton, UK, whose hospitality and financial support is gratefully acknowledged. We thank two anonymous Referees and an Associate Editor for helping us improve and extend the scope of our paper.

under \mathcal{M}_i , $i = 0, 1$. Then, the BF in favor of \mathcal{M}_1 , or against \mathcal{M}_0 , is defined as $BF_{10}(y) = m_1(y)/m_0(y)$, where $m_i(y) = \int f_i(y|\theta_i)p_i(\theta_i)d\theta_i$ is the marginal density of y under \mathcal{M}_i , also called the marginal likelihood of \mathcal{M}_i .

It is well known that special care must be exercised in the specification of $p_0(\cdot)$ and $p_1(\cdot)$ when computing the BF. One obvious condition is that neither prior be improper, because the resulting BF would depend on arbitrary constants. Even when proper priors are used, however, difficulties may arise; in particular, this happens when $p_1(\cdot)$ is not chosen in view of the comparison with model \mathcal{M}_0 , which is of course the rule with conventional priors. In general, a conventional $p_1(\cdot)$ will be rather diffuse, and will thus give little weight to sampling densities close to the subspace characterizing \mathcal{M}_0 . Therefore, unless the data are vastly against \mathcal{M}_0 , which rarely happens for moderate sample sizes, there will be an evidence bias in favor of \mathcal{M}_0 . Informally, this happens because $p_1(\cdot)$ “wastes” probability mass in parameter areas too remote from the null. This fact had essentially been realized as early as in Jeffreys (1961, Ch. 3) and was already clear in Morris (1987), whose suggestion was to “center” $p_1(\cdot)$ around the null-subspace. In this spirit, we will argue in favor of “transferring probability mass” towards the null subspace within a given diffuse prior under \mathcal{M}_1 .

Although we used no limiting argument above, there is a connection with the Jeffreys-Lindley-Bartlett paradox; see O’Hagan and Forster (2004, Sect. 3.33) and Robert, Chopin and Rousseau (2009). According to one version of the paradox, if the sample size is fixed, but the variance of $p_1(\cdot)$ is free to increase without bound, the posterior probability of the null model will go to one, irrespective of the data. This is just an exacerbation of the phenomenon described above, with $p_1(\cdot)$ allocating probability mass in unreasonable regions of the parameter space.

A word of caution is useful at this stage. From a Bayesian perspective, a model is a *pair*, whose elements are the family of sampling distributions (sampling model) and the prior. Nevertheless, we will follow the prevailing practice of using the word “model” to identify the sampling model, leaving to the prior the role of specifying which Bayesian model is actually entertained.

Adhering to an objective viewpoint, we assume that default parameter priors $p_i(\cdot)$, $i = 1, 2$, are given, each of them depending only on the corresponding model. We also assume, for simplicity, that both priors are proper. The action of reallocating mass within $p_1(\cdot)$ towards the null subspace has a negative side effect, at least for moderate sample sizes: it will diminish evidence in favor of \mathcal{M}_1 when the parameter values generating the data are truly away from the null. However, this price is worth paying, to some extent, because of two reasons: i) the very fact that we are considering \mathcal{M}_0 testifies that it has some *a priori* plausibility and thus parameter values close to the null are more interesting to monitor than those remote from it; ii) if the data manifestly support \mathcal{M}_1 , we can surely afford the luxury to somewhat diminish the strength of evidence in its favor, because it will be already high enough for most practical purposes. However, it is not at all obvious how far this strategy should be pushed, and we dedicate part of this paper to try and answer this question.

In light of the above discussion, two general issues are to be addressed in the setting under consideration:

- 1) given model \mathcal{M}_0 nested in \mathcal{M}_1 , and the corresponding default priors $p_0(\cdot)$ and $p_1(\cdot)$, how can we build an \mathcal{M}_0 -focussed prior under \mathcal{M}_1 , transferring

- probability mass within $p_1(\cdot)$ toward the null subspace characterizing \mathcal{M}_0 ?
- 2) how do we settle the *evidence trade-off*: reinforcing the evidence in favor of \mathcal{M}_1 for parameter values around the null subspace, while weakening it when the parameter lies in regions away from the null?

A possible objection is that point 1) could be bypassed: once we have understood the features of a “good” parameter prior under \mathcal{M}_1 , why should we bother with the default prior anyway? We accept this criticism, and do not object to a subjective specification carefully taking into account the *desiderata* we set out. We remark however that this task may be far from simple, requires substantive knowledge not always available, and could become daunting when many pairwise comparisons are entertained (like in variable selection). This is the reason why we privilege an objective approach, which takes as input only the default priors.

A natural answer to point 1) is provided by the *intrinsic priors*, whose scope is indeed not restricted to nested models. Intrinsic priors are now recognized as an important tool for objective Bayesian hypothesis testing and model comparison. Numerous applications witness their usefulness, ranging from variable selection (Casella and Moreno, 2006; Casella et al., 2009; Moreno, Girón and Casella, 2010; Leon-Novelo, Moreno and Casella, 2012) to contingency tables (Casella and Moreno, 2005; Consonni and La Rocca, 2008; Casella and Moreno, 2009; Consonni, Moreno and Venturini, 2011) to change point problems (Moreno, Casella and Garcia-Ferrer, 2005; Girón, Moreno and Casella, 2007). When the two models are nested, the end result of the intrinsic prior procedure is to modify $p_1(\cdot)$ so that the resulting intrinsic prior accumulates more mass around the null subspace. This is achieved by mixing over a *training sample*, whose size t regulates the amount of concentration of the intrinsic prior, which we denote by $p_1^I(\cdot|t)$, around the null subspace.

If $p_1(\cdot)$ is improper, it is tempting to set t in the intrinsic prior $p_1^I(\cdot|t)$ equal to the *minimal* training sample size, which is the smallest sample size for which the default posterior is proper on all data. However, no formal justification is available for this choice, which clearly bypasses point 2) on grounds of simplicity. On the other hand, when the default prior is proper, as happens in some discrete data problems, there is no general guideline for fixing t , and usually a robustness analysis is performed by letting t vary between 1 and n , where n is the actual sample size; see for instance Casella and Moreno (2009). Point 2) is here bypassed in favor of a sensitivity analysis.

Intrinsic priors for the comparison of nested models can be viewed as *expected posterior priors* (Pérez and Berger, 2002) with baseline mixing distribution equal to the marginal data distribution under \mathcal{M}_0 . Another related approach is due to Neal (2001): since subjective prior elicitation of the parameter prior should be more precise, and possibly easier, under the smaller model than under the larger model, information should be transferred from the former to the latter by means of a training sample, whose sample size t will determine how similar or compatible the two models turn out to be. Neal offers no guidance on fixing t ; interestingly, however, in his approach t can grow to infinity.

An alternative to the intrinsic (or expected posterior) prior approach to derive the BF in the presence of improper priors is the Fractional Bayes Factor (FBF); see O’Hagan (1995). Here a fraction b of the likelihood is used to obtain a fractional posterior distribution, which in turn is used as a (data-dependent)

prior to construct a fractional marginal likelihood based on the likelihood raised to the complementary fraction $(1 - b)$. This calculation is repeated under both models. The end result is to shift the prior under each of the two models towards a region supported by the likelihood. Clearly, if the data are in reasonable accord with \mathcal{M}_0 , which we have identified as the most critical situation for nested model comparison, the prior under \mathcal{M}_1 will tend to concentrate around the null subspace, like in the intrinsic prior approach. In this sense, the fraction b plays a role akin to that of the training sample size t , although it should be stressed that the implied prior in the FBF is data dependent, while this is not the case for the intrinsic prior. A conventional choice is $b = n_0/n$, where n_0 is the smallest integer that makes the fractional posteriors proper. O'Hagan (1995, Sect. 6) also suggests two alternative choices for cases when robustness is a major concern, but Moreno (1997) has an argument against these choices. Data-centered priors for each of the two models can also be constructed using the expected posterior priors of Pérez and Berger (2002), setting the baseline mixing distribution equal to the empirical distribution.

The FBF is typically easier to implement than the intrinsic approach. However, the fact that it uses a data dependent prior is clearly a drawback. Accordingly, since implementation issues turn out to be less compelling for discrete data problems, in this paper we address point 1) through intrinsic priors. As for the issue raised in point 2), which to the best of our knowledge has never been tackled so far, we propose a solution in Subsection 3.1 based on the notion of *total weight of evidence*. We do not claim that our solution is universal, but we found it useful in some examples, and we believe that it sheds light on the evidence trade-off.

We now turn to a related aspect, which has been relatively neglected in the literature on priors for model comparison: the asymmetry in the learning rate of the BF between two nested models \mathcal{M}_0 and \mathcal{M}_1 . A typical prior $p_1(\cdot)$, whether subjective or objective, is continuous and strictly positive on the null subspace. The second condition (given the first one) makes it a *local* prior. A serious deficiency of local priors relates to their asymptotic learning rate. Specifically, the BF in favor of \mathcal{M}_1 , when \mathcal{M}_1 holds, diverges in probability exponentially fast, as the sample size grows; on the other hand, when \mathcal{M}_0 holds, the same BF converges to zero in probability at a polynomial rate only. Although this fact is well established, it did not receive very much attention until Johnson and Rossell (2010) brought it to the fore. Robert, Chopin and Rousseau (2009, Sect. 7) report evidence that Jeffreys was aware of the asymmetry, but that later studies neglected it. In practice, the problem is that the imbalance is already quite dramatic for moderate sample sizes. However, as suggested by Johnson and Rossell (2010), this unsatisfactory feature can be corrected by using *non-local* priors. As the name suggests, these priors are built in opposition to local priors, and their distinguishing feature, assuming continuity, is to be identically zero on the null subspace.

We find the idea of non-local priors appealing, not only because they ameliorate the learning rate of the BF, but also because they force the user to think more carefully about the notion of *model separation*. This is a difficult issue, of course, which only occasionally can be answered employing subject-matter knowledge; a notable example, reported in Cohen (1992), is that standardized effect sizes of less than 0.2, in absolute value, are often not considered substantively important

in the social sciences. However, we hasten to say that non-local priors have been disapproved by some authors; see for instance the discussions of Consonni and La Rocca (2011) by J. Q. Smith and J. Rousseau with C. P. Robert.

Intrinsic priors and non-local priors play complementary roles in the comparison of nested models. If the BF is seen as an implementation of Ockham's razor, the principle that an explanation (model) should not be more complicated than necessary (*pluralitas non est ponenda sine necessitate*), as suggested by Jaynes (1979), Smith and Spiegelhalter (1980), and Jefferys and Berger (1992), then Bayesian barbers should worry both about the sharpness of their tool, on the one hand, and the risk of cutting the throat of the larger model, on the other hand. Intrinsic priors protect the larger model from being treated unfairly, and thus play the role of an alum block, whereas non-local priors can greatly sharpen the blade of the razor, and thus play the role of a whetstone. A skilled combination of the two tools helps the Bayesian barber to achieve a balanced comparison of the two models. In fact, we show in this paper that a suitably defined family of non-local intrinsic priors produces a BF with finite sample properties comparable to those of ordinary (local) intrinsic priors, and with the improved learning rate (when the null model holds) characterizing non-local priors.

The rest of the paper is organized as follows. Section 2 provides background material on intrinsic priors and on a particular class of non-local priors, *moment priors*, using as illustration the problem of testing a sharp null hypothesis on a Bernoulli proportion. Section 3 presents the class of *intrinsic moment priors*, for the comparison of two nested models, which is implemented in Section 4 for testing the equality of two proportions and in Section 5 for variable selection in logistic regression models. Section 6 applies the suggested testing procedures to a collection of randomized binary trials of a new surgical treatment for stomach ulcers, also discussed from a meta-analysis perspective by Efron (1996), and to a medical data set already analyzed by Dellaportas, Forster and Ntzoufras (2002) using logistic regression models. Finally, Section 7 offers some concluding remarks and investigates a few issues worth of further consideration. A technical Appendix on the asymptotic learning rate of BFs completes the paper.

2. PRIORS FOR THE COMPARISON OF NESTED MODELS

We review in this section two methodologies for constructing priors when two nested models are compared: intrinsic priors and a specific class of non-local priors, called moment priors.

Consider two sampling models for the same *discrete* vector of observables y :

$$(1) \quad \mathcal{M}_0 = \{f_0(\cdot|\xi_0), \xi_0 \in \Xi_0\} \quad \text{vs} \quad \mathcal{M}_1 = \{f_1(\cdot|\xi_1), \xi_1 \in \Xi_1\},$$

where \mathcal{M}_0 is nested in \mathcal{M}_1 , i.e., for all $\xi_0 \in \Xi_0$, $f_0(\cdot|\xi_0) = f_1(\cdot|\xi_1)$, for some $\xi_1 \in \tilde{\Xi}_0 \subset \Xi_1$, where $\tilde{\Xi}_0$ is isomorphic to Ξ_0 and of lower dimensionality than Ξ_1 . Let $p_0(\cdot)$ be a given prior under \mathcal{M}_0 , and similarly for $p_1(\cdot)$ under \mathcal{M}_1 , both of them being *proper*; this assumption simplifies the exposition and is not particularly restrictive because we deal with discrete data models. Typically both $p_0(\cdot)$ and $p_1(\cdot)$ will be default, inference-based, priors. We also assume, again for simplicity, equal prior probabilities for \mathcal{M}_0 and \mathcal{M}_1 , so that the posterior probability of \mathcal{M}_1 is a function of the BF only: $\mathbb{P}(\mathcal{M}_1|y) = (1 + BF_{01}(y))^{-1}$, where $BF_{01}(y) = 1/BF_{10}(y)$ is the BF in favor of \mathcal{M}_0 .

2.1 Intrinsic priors

Intrinsic priors were introduced in objective hypothesis testing to deal meaningfully with improper default priors when constructing BFs; see Berger and Pericchi (1996); Moreno (1997); Moreno, Bertolino and Racugno (1998). However, this view of the intrinsic prior approach is unduly restrictive and actually hinders its inherent nature, as it is apparent for discrete data models: in this case default priors are usually proper, but the intrinsic approach can still be very useful.

As recalled in the Introduction, a default prior $p_1(\cdot)$ is typically inappropriate for testing purposes, because it assigns little mass around the null subspace $\tilde{\Xi}_0$. Mixing over the training sample $x = (x_1, \dots, x_t)$, the intrinsic prior on ξ_1 can be written as

$$(2) \quad p_1^I(\xi_1|t) = \sum_x p_1(\xi_1|x)m_0(x), \quad \xi_1 \in \Xi_1,$$

where $p_1(\xi_1|x)$ is the posterior density of ξ_1 under \mathcal{M}_1 , given x , and $m_0(x) = \int f_0(x|\xi_0)p_0(\xi_0)d\xi_0$ is the marginal density of x under \mathcal{M}_0 ; it is natural to let $t = 0$ in $p_1^I(\cdot|t)$ return the default prior $p_1(\cdot)$.

We remark that (2) is not the original definition of intrinsic prior, but rather its formulation as an expected posterior prior (Pérez and Berger, 2002). We find formula (2) especially appealing, because it makes clear that an intrinsic prior is a mixture of “posterior” distributions. As we will illustrate shortly, if the training sample size t grows, the intrinsic prior increases its concentration on the subspace $\tilde{\Xi}_0$. This is apparent from (2), because the weights $m_0(x)$ in the mixture will be higher for realizations x more likely under \mathcal{M}_0 , and these realizations x will drive the posterior $p_1(\cdot|x)$ towards parameter values more supported under \mathcal{M}_0 . Notice that, if t grows to infinity, the two Bayesian models ($\{f_0(\cdot|\xi_0), \xi_0 \in \Xi_0\}, p_0(\cdot)$) and ($\{f_1(\cdot|\xi_1), \xi_1 \in \Xi_1\}, p_1(\cdot)$) will coincide, making the comparison problem trivial. Subsection 3.1 will discuss in greater detail the nature of t , and will present a method to choose its value.

The BF based on the intrinsic prior is a weighted average of conditional BFs based on the default prior:

$$(3) \quad BF_{10}^I(y|t) = \sum_x BF_{10}(y|x)m_0(x),$$

where $BF_{10}(y|x)$ is the BF obtained using $p_1(\cdot|x)$ as prior under model \mathcal{M}_1 ; see for example Consonni and La Rocca (2008, Proposition 3.4). Hence, at least for small t and conjugate $p_1(\cdot)$, computing $BF_{10}^I(y|t)$ is not much more demanding than computing $BF_{10}(y)$.

EXAMPLE 2.1 (Bernoulli). *Consider the testing problem $\mathcal{M}_0 : f_0(y|\theta_0) = \text{Bin}(y|n, \theta_0)$ versus $\mathcal{M}_1 : f_1(y|\theta) = \text{Bin}(y|n, \theta)$, where θ_0 is a fixed value, while θ varies in $(0, 1)$. Let the default prior be $p_1(\theta|b) = \text{Beta}(\theta|b, b)$ for some $b > 0$. We take a symmetric prior because default objective priors typically satisfy this property. In particular, letting $b = 1/2$ we obtain Jeffreys’s prior, whereas $b = 1$ gives us the uniform prior. The intrinsic prior in this example is given by*

$$(4) \quad p_1^I(\theta|b, t) = \sum_{x=0}^t \text{Beta}(\theta|b+x, b+t-x)\text{Bin}(x|t, \theta_0).$$

The solid curves in Figure 1(a), i.e., those specified by $h = 0$, illustrate the shape of the intrinsic priors with training sample size $t = 0$ (default prior), $t = 1$ and $t = 8$, when $\theta_0 = 0.25$ and $b = 1$. The dashed curves ($h = 1$) should be disregarded for the time being. The effect of the intrinsic procedure is very clear: already with $t = 1$ the density has become a straight line with negative slope, so as to start privileging low values of θ , such as $\theta_0 = 0.25$, and with a training sample size $t = 8$ the effect is much more dramatic, with the density now having a mode somewhere around 0.25 and then declining quickly.

Figure 1(b) shows (again focus on solid lines only) the effect of the above-described probability mass transfer on the comparison between \mathcal{M}_0 and \mathcal{M}_1 : for a small sample situation ($n = 12$) the posterior probability of \mathcal{M}_1 , computed from

$$(5) \quad BF_{10}^I(y|b, t) = \sum_{x=0}^t \frac{B(b+x+y, b+t-x+n-y)}{B(b+x, b+t-x)\theta_0^y(1-\theta_0)^{n-y}} \text{Bin}(x|t, \theta_0),$$

where $B(\cdot, \cdot)$ denotes the Beta function, is represented as a function of the observed frequency \bar{y} (evidence curve) both for the default prior and for the intrinsic prior with $t = 1$. The evidence curve reaches a minimum at $\bar{y} = 0.25$ (data perfectly supporting the null) and is somewhat higher for the intrinsic prior than for the default prior when $0 < \bar{y} < 0.5$, because with the intrinsic prior \mathcal{M}_1 becomes a stronger competitor when the data moderately support \mathcal{M}_0 .

Results similar to those given by the intrinsic prior can be obtained, in the above example, by using a suitable Beta prior centered at θ_0 . However, the probability mass transfer towards θ_0 takes place more smoothly under the intrinsic prior than under a Beta prior with increasing precision, because the intrinsic prior is a mixture of Beta distributions. Moreover, this kind of alternative approach is only available because we are testing a sharp hypothesis. When testing a composite hypothesis, it is not at all obvious that a suitable conjugate prior can be found (after a reparametrization of the model to identify a parameter of interest and a nuisance parameter). On the other hand, as we will see in Section 4 for the comparison of two proportions, the intrinsic prior produces the desired outcome in a natural and automatic way.

2.2 Moment Priors

Consider the testing problem (1). We say that *the smaller model holds* if the sampling distribution of the data belongs to \mathcal{M}_0 ; we say that *the larger model holds* if it belongs to \mathcal{M}_1 but not to \mathcal{M}_0 . The following result shows an imbalance in the learning rate of the BF for commonly used priors.

RESULT 2.1. *In the testing problem (1) assume that $p_0(\cdot)$ and $p_1(\cdot)$ are continuous and strictly positive on Ξ_0 and Ξ_1 , respectively, that some regularity conditions are satisfied by the two models, and that the data $y^{(n)} = (y_1, \dots, y_n)$ arise under i.i.d. sampling. If \mathcal{M}_0 holds, then $BF_{10}(y^{(n)}) = n^{-(d_1-d_0)/2} e^{O_p(1)}$, as $n \rightarrow \infty$, where d_j is the dimension of Ξ_j , $j = 1, 2$, with $d_1 > d_0$; if \mathcal{M}_1 holds, then $BF_{01}(y^{(n)}) = e^{-Kn+O_p(n^{1/2})}$, as $n \rightarrow \infty$, for some $K > 0$.*

We refer to Dawid (2011) for a proof of this result. It should be noted that a crucial role is played by the fact that $p_1(\xi_1) > 0$ for all $\xi_1 \in \tilde{\Xi}_0$; also recall that

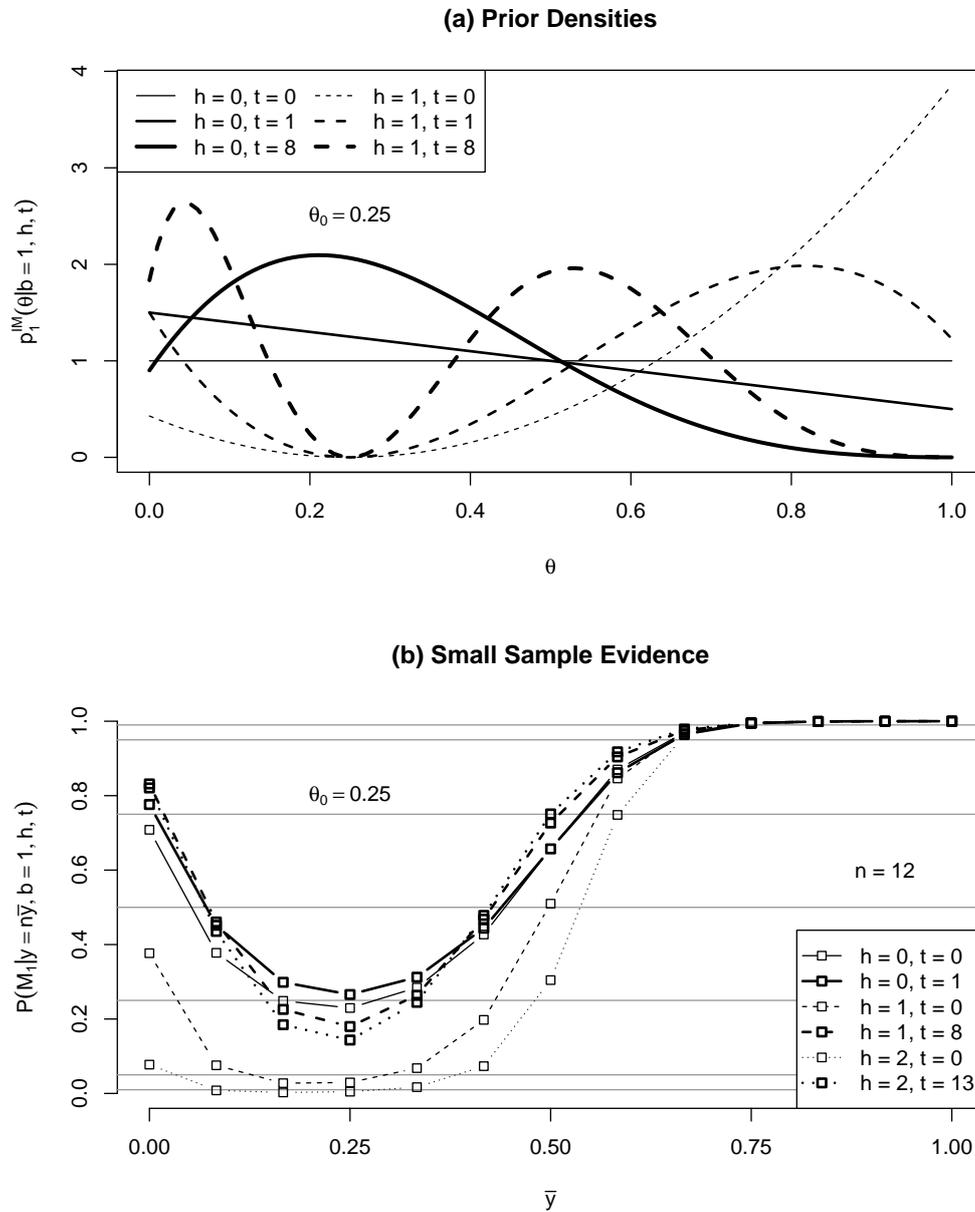


FIG 1. Prior densities (a) and small sample evidence (b) for the Bernoulli example. Horizontal gray lines in (b) denote possible decision thresholds at 1%, 5%, 25%, 50%, 75%, 95% and 99% on the posterior probability scale.

$p_1(\xi_1)$ is continuous. Thus the only way to speed up the decrease of $BF_{10}(y^{(n)})$, when \mathcal{M}_0 holds, is to force the prior density under \mathcal{M}_1 to vanish on $\tilde{\Xi}_0$.

Let $g_h(\cdot)$ be a smooth function from Ξ_1 to \mathfrak{R}_+ vanishing on $\tilde{\Xi}_0$, together with its first $2h - 1$ derivatives, while $g_h^{(2h)}(\xi)$ is different from zero for all $\xi \in \tilde{\Xi}_0$; assume that $\int_{\Xi_1} g_h(\xi_1)p_1(\xi_1)d\xi_1$ is finite and non-zero. Starting from a given local prior $p_1(\cdot)$, we define the *generalized* moment prior with moment function $g_h(\cdot)$ as

$$(6) \quad p_1^M(\xi_1|h) \propto g_h(\xi_1)p_1(\xi_1), \quad \xi_1 \in \Xi_1.$$

We impose that $g_0(\xi_1) \equiv 1$, so that setting $h = 0$ in $p_1^M(\cdot|h)$ returns the local prior $p_1(\cdot)$. For instance, if $\Xi_1 \subseteq \mathfrak{R}$ and $\tilde{\Xi}_0 = \Xi_0 = \{\xi_0\}$, with ξ_0 a fixed value, we may take $g_h(\xi_1) = (\xi_1 - \xi_0)^{2h}$; this defines the moment prior introduced by Johnson and Rossell (2010) for testing a sharp hypothesis on a scalar parameter. We refer to h as the *order* of the (generalized) moment prior.

The BF against \mathcal{M}_0 based on prior (6) can be computed as

$$(7) \quad BF_{10}^M(y^{(n)}|h) = \frac{\int_{\Xi_1} g_h(\xi_1)p_1(\xi_1|y^{(n)})d\xi_1}{\int_{\Xi_1} g_h(\xi_1)p_1(\xi_1)d\xi_1} BF_{10}(y^{(n)}),$$

so that the extra effort required by using this prior amounts to computing some (generalized) moments of the local prior and posterior. This effort is rewarded by a reduction in the learning rate imbalance: $BF_{10}^M(y^{(n)}|h) = n^{-h-(d_1-d_0)/2}e^{O_p(1)}$, when \mathcal{M}_0 holds, while we still have $BF_{01}^M(y^{(n)}|h) = e^{-Kn+O_p(n^{1/2})}$, when \mathcal{M}_1 holds; see the Appendix for a justification of this result, which generalizes the rates found by Johnson and Rossell (2010) for their specific moment priors.

EXAMPLE 2.2 (Bernoulli ctd). *Starting from the local prior $\text{Beta}(\theta|a_1, a_2)$, we define the moment prior of order h as*

$$(8) \quad p_1^M(\theta|a_1, a_2, h) = \frac{(\theta - \theta_0)^{2h}}{K(a_1, a_2, h, \theta_0)} \text{Beta}(\theta|a_1, a_2),$$

where

$$(9) \quad K(a_1, a_2, h, \theta_0) = \frac{\theta_0^{2h}}{B(a_1, a_2)} \sum_{j=0}^{2h} \binom{2h}{j} (-1)^j \theta_0^{-j} B(a_1 + j, a_2),$$

and obtain

$$(10) \quad BF_{10}^M(y|a_1, a_2, h) = \frac{K(a_1 + y, a_2 + n - y, h, \theta_0)}{K(a_1, a_2, h, \theta_0)} \frac{B(a_1 + y, a_2 + n - y)}{B(a_1, a_2)\theta_0^y(1 - \theta_0)^{n-y}}.$$

In particular, we are interested in the default choice $a_1 = a_2 = b$ (with $b = 1/2$ or $b = 1$). The moment prior $p_1^M(\theta|b, h)$ is represented in Figure 1(a), for $b = 1$ and $\theta_0 = 0.25$, by the thin dashed curve specified by $h = 1$ and $t = 0$. The other two dashed curves, specified by $h = 1$ and $t = 1$ (intermediate curve) or $t = 8$ (thick curve) should be ignored for the time being. The shape of $p_1^M(\theta|b, h)$ can be described as follows: it is zero at the null value $\theta_0 = 0.25$, as required, it increases rapidly as θ goes to 1, while it goes up more gently as θ goes to zero. It is clear that this moment prior will not be suitable for testing purposes, because it puts too

much mass away from θ_0 . This is confirmed by the thin dashed line in Figure 1(b): the null model is unduly favored. The thin dotted line in the same plot shows that things get even worse for $h = 2$ and $t = 0$; again, for now, please disregard the curves with $t = 8$ and $t = 13$. On the other hand, the moment prior has an improved learning rate as the sample size grows: we postpone the illustration of this feature to the next section, after the moment prior has been made suitable for testing purposes by means of a probability mass transfer towards the null value θ_0 .

Rousseau and Robert, in discussing Consonni and La Rocca (2011), raise an interesting point in relation to moment priors. They cast the problem in a decision-theoretic setup and use the well-known duality between prior and loss function (Rubin, 1987; Robert, 2001) to suggest that non-local priors should be replaced by the use of suitable loss functions, which take into account the distance from the null. This perspective was actually pursued in Robert and Casella (1994); see also Goutis and Robert (1998). Indeed, it can be checked that the optimal Bayesian decision, under a $\{0,1\}$ -loss function and a moment prior of the form $p_1^M(\xi_1) \propto (\xi_1 - \xi_0)^{2h} p_1(\xi_1)$, where $p_1(\xi_1)$ is a local prior, and ξ_0 a null parameter value, coincides with that arising from the local prior $p_1(\xi_1)$ and a “distance weighted” loss function of the form

$$L(a, \xi) = \begin{cases} K_1 & \text{if } a = 1 \quad \& \quad \xi = \xi_0 \\ (\xi - \xi_0)^{2h} & \text{if } a = 0 \quad \& \quad \xi \neq \xi_0 \\ 0 & \text{otherwise,} \end{cases}$$

where a is the action, taking value 0 or 1 if the chosen model is \mathcal{M}_0 or \mathcal{M}_1 , respectively, while $K_1 = \mathbb{E}_1[(\xi_1 - \xi_0)^{2h}]$ is the expected loss, under the local prior $p_1(\xi_1)$, when \mathcal{M}_0 is wrongly chosen. This interpretation of moment priors is interesting and, from our viewpoint, it reinforces their usefulness, because it shows that a moment prior can be justified using decision theory.

3. INTRINSIC MOMENT PRIORS

Example 2.2 shows that the moment prior obtained from a default local prior, call it the *default moment prior*, does not accumulate enough mass around the null value (more generally around the subspace specified by the null model). This suggests applying the intrinsic procedure to the default moment prior, obtaining in this way a new class of priors for testing nested hypotheses, which we name intrinsic moment priors. The improved learning rate extends to the latter priors, because each of them is a mixture (through the intrinsic procedure) of non-local priors.

Our strategy for a balanced objective Bayesian comparison of two nested models thus starts with a default prior under each of the two models and then envisages two steps: i) construct the default moment prior of order h under the larger model; ii) for a given training sample size t , generate the corresponding intrinsic moment prior. We recommend using the resulting prior to compute the BF: step i) improves the learning rate (when the null model holds), while step ii) makes sure that the testing procedure exhibits a good small sample behavior in terms of the evidence curve. We first illustrate intrinsic moment priors in our running example, then we discuss the choice of t . In Section 4 the procedure will be implemented to test the equality of two proportions, while in Section 5 it will be developed for the family of logistic regression models.

EXAMPLE 3.1 (Bernoulli ctd). Recall that the intrinsic prior is an average of “posterior” distributions. Since in our case we start from the moment prior (8) with default choice $a_1 = a_2 = b$, the intrinsic moment prior for θ with training sample size t will be given by

$$(11) \quad p_1^{IM}(\theta|b, h, t) = \sum_{x=0}^t \frac{(\theta - \theta_0)^{2h} \text{Beta}(\theta|b+x, b+t-x)}{K(b+x, b+t-x, h, \theta_0)} \text{Bin}(x|t, \theta_0),$$

where $K(\cdot, \cdot, h, \theta_0)$ is defined in (9), and we exploited conjugacy. Notice that (11) describes a family of prior distributions including the standard intrinsic prior ($h = 0$) and the default prior ($h = 0, t = 0$) as special cases. Similarly, from (3) we find

$$(12) \quad BF_{10}^{IM}(y|b, h, t) = \sum_{x=0}^t BF_{10}^M(y|b+x, b+t-x, h) \text{Bin}(x|t, \theta_0),$$

where $BF_{10}^M(y|\cdot, \cdot, h)$ is defined in (10).

Figure 1(a) shows (letting $b = 1$) the effect of applying the intrinsic procedure to the default moment prior of order $h = 1$ (dashed curves): as t grows, the overall shape of the prior density changes considerably, because more and more probability mass in the extremes is displaced towards θ_0 , giving rise to two modes, while the non-local nature of the prior is preserved, because the density remains zero at $\theta_0 = 0.25$. In this way, as shown in Figure 1(b), the evidence against the null for small samples is brought back to more reasonable values (with respect to the default moment prior). More specifically, Figure 1(b) shows that the intrinsic moment prior with $h = 1$ and $t = 8$ (a choice explained later in Subsection 3.1) performs comparably to the uniform prior (and to the standard intrinsic prior with unit training sample) over a broad range of values for the observed sampling fraction \bar{y} ; this intrinsic moment prior results in a smaller amount of evidence (against \mathcal{M}_0) for values of \bar{y} close to $0.25 = \theta_0$, which is to be expected for continuity, but induces a steeper evidence gradient as \bar{y} moves away from the null point in either direction, which makes it appealing.

The learning rate of the intrinsic moment prior is illustrated in Figure 2(a), which reports the average posterior probability of the null model when $\theta = 0.25$ (null value) and when $\theta = 0.4$ (an instance of the alternative model). It is apparent from this plot that a non-local prior ($h > 0$) is needed, if strong evidence in favor of the null has “ever” to be achieved, but also that the intrinsic procedure is crucial to calibrate small sample evidence. These results are striking, and they signal that our strategy actually represents a marked improvement over current methods. Notice that there is an associated cost: the moment prior trades off a delay in learning the alternative model for speed in learning the null model; the intrinsic procedure is remarkably effective in controlling this trade off. In light of Figure 2(a), we recommend letting $h = 1$ by default, and trying $h = 2$ for sensitivity purposes; we remark that $h = 1$ is enough to change the convergence rate of $BF_{10}^{IM}(y^{(n)})$, when \mathcal{M}_0 holds, from sub-linear to super-linear.

3.1 Choosing the training sample size

Recall that the goal of the intrinsic procedure is to transfer probability mass toward the null subspace within the default prior under \mathcal{M}_1 . There is clearly a

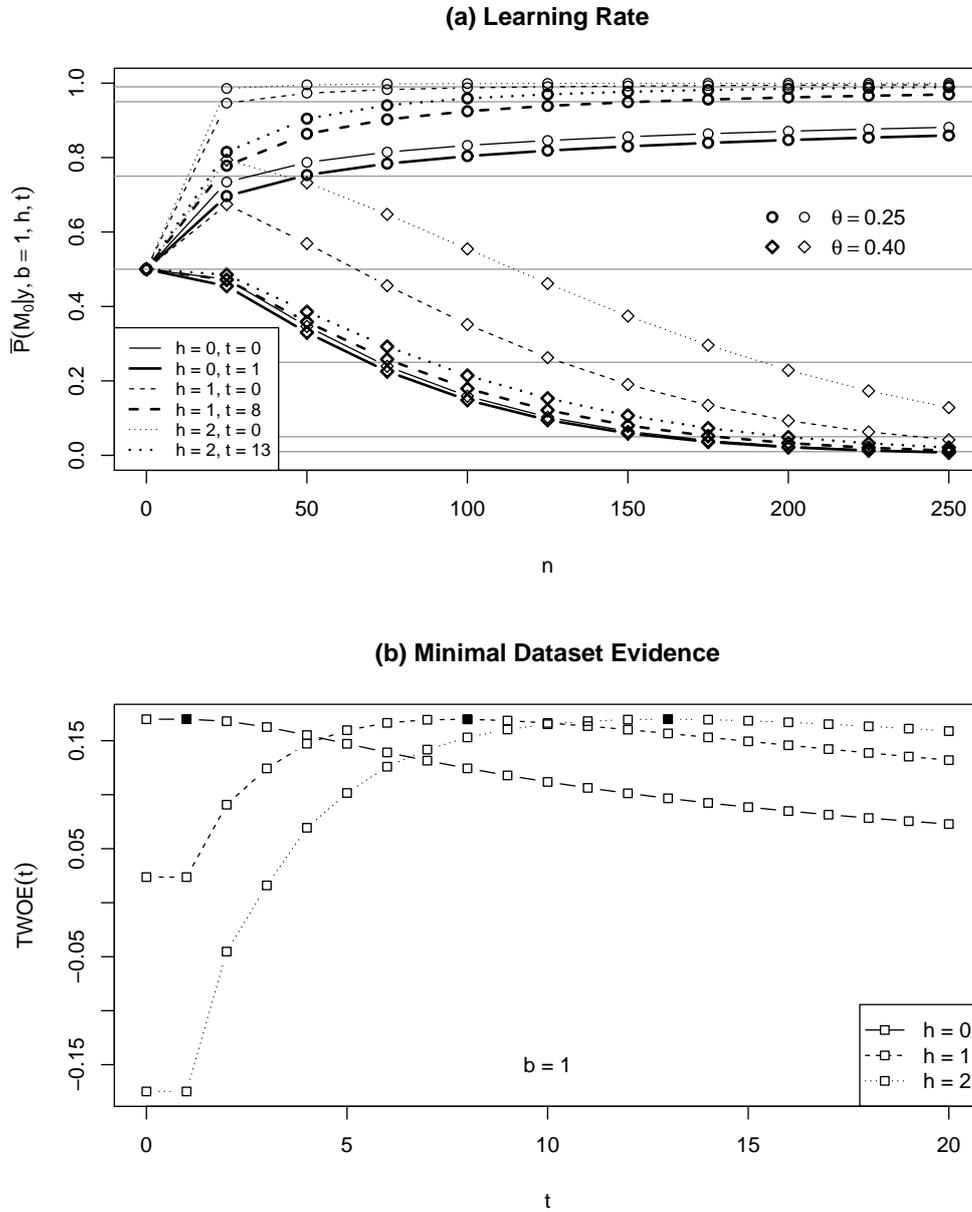


FIG 2. Learning rate (a) and minimal data set evidence (b) for the Bernoulli example. Horizontal gray lines in (a) denote possible decision thresholds at 1%, 5%, 25%, 50%, 75%, 95% and 99% on the posterior probability scale.

tension here between this aim and that of leaving enough mass in other areas of the parameter space, not to unduly discredit \mathcal{M}_1 . This is precisely the issue we face when choosing t . We now provide some guidelines for the Bernoulli problem, with a view to more general situations.

We aim at a single recommended value of t for all possible values of θ_0 . For this purpose, we fix $\theta_0 = 1/2$, representing the worst case scenario in terms of the information content of a single observation. A minimal sample size to discriminate between \mathcal{M}_1 and \mathcal{M}_0 is $n = 2$, with possible data values $y = 0, 1, 2$. Define the *weight of evidence* against the null using an intrinsic moment prior as $WOE_y(t) = \log BF_{10}^{IM}(y|b, h, t)$, where we focus on the dependence on t for a given choice of b and h . By symmetry $WOE_0(t) = WOE_2(t)$. However $WOE_1(t) \neq WOE_0(t)$; this is why we are able to discriminate between the two models if $n = 2$, which would not happen with $n = 1$. It can be checked that $WOE_1(t)$ is increasing in t , while $WOE_0(t) = WOE_2(t)$ is decreasing in t . The explanation of this phenomenon is simple, keeping in mind that an increase in t transfers probability mass towards $\theta_0 = 1/2$ within the prior: the value $y = 1$ supports \mathcal{M}_0 , and thus its marginal probability under \mathcal{M}_1 will increase with t ; the values $y = 0$ and $y = 2$ support \mathcal{M}_1 , and thus an increase in t will make their marginal probability under \mathcal{M}_1 smaller. How far should we let t grow? To answer this question, define the *total weight of evidence* $TWOE(t) = \sum_y WOE_y(t)$, and consider the weight of evidence as a sort of currency: we will be willing to trade off a decrease in $WOE_0(t)$ and $WOE_2(t)$ for an increase in $WOE_1(t)$ as long as we get more than we give, that is, as long as we increase $TWOE(t)$. Define $t^* = \operatorname{argmax}_t TWOE(t)$, and assume that this quantity is well-defined. The value t^* represents our optimal training sample size when implementing the intrinsic procedure. Since this choice of t is based on a somewhat unusual criterion, we will be willing to let t vary in a neighbourhood of t^* for a sensitivity analysis.

We remark that the above strategy to find t^* in an intrinsic procedure is general, at least for finitely discrete data models. In particular, it can be used to determine an optimal training sample size also for the standard intrinsic prior ($h = 0$). Figure 2(b) plots $TWOE(t)$ for $h = 0, 1, 2$, assuming a uniform default prior ($b = 1$). Interestingly, when $h = 0$ (standard intrinsic prior), we find $t^* \in \{0, 1\}$. This seeming indeterminacy can be explained by noticing that, when $\theta_0 = 0.5$, the intrinsic prior with $t = 1$ is the uniform prior, i.e., it is the same as the default prior (corresponding to $t = 0$). On the other hand, when the starting prior is the default moment prior of order $h = 1$, it turns out that $t^* = 8$, while for $h = 2$ we obtain $t^* = 13$, so that with non-local moment priors the intrinsic procedure is necessary: this makes sense, because the starting prior puts mass at the endpoints of the parameter space in a rather extreme way.

4. TESTING THE EQUALITY OF TWO PROPORTIONS

Suppose the larger (encompassing) model is the product of two binomial models

$$(13) \quad \mathcal{M}_1 : f_1(y_1, y_2 | \theta_1, \theta_2) = \operatorname{Bin}(y_1 | n_1, \theta_1) \operatorname{Bin}(y_2 | n_2, \theta_2),$$

where n_1 and n_2 are fixed sample sizes. The null model assumes $\theta_1 = \theta_2 = \theta$, so that

$$(14) \quad \mathcal{M}_0 : f_0(y_1, y_2 | \theta) = \operatorname{Bin}(y_1 | n_1, \theta) \operatorname{Bin}(y_2 | n_2, \theta).$$

A default prior for θ under \mathcal{M}_0 is $p_0(\theta|b_0) = \text{Beta}(\theta|b_0, b_0)$, while a default prior for (θ_1, θ_2) under \mathcal{M}_1 is given by $p_1(\theta_1, \theta_2|b_1, b_2) = \text{Beta}(\theta_1|b_1, b_1)\text{Beta}(\theta_2|b_2, b_2)$.

Starting from a more general conjugate prior $\text{Beta}(\theta_1|a_{11}, a_{12})\text{Beta}(\theta_2|a_{21}, a_{22})$ under \mathcal{M}_1 , we define the moment prior of order h as

$$(15) \quad p_1^M(\theta_1, \theta_2|a, h) = \frac{(\theta_1 - \theta_2)^{2h}}{K(a, h)} \text{Beta}(\theta_1|a_{11}, a_{12})\text{Beta}(\theta_2|a_{21}, a_{22}),$$

where $a = [[a_{jk}]_{k=1,2}]_{j=1,2}$ is a matrix of strictly positive real numbers and

$$(16) \quad K(a, h) = \sum_{j=0}^{2h} \binom{2h}{j} (-1)^j \frac{B(a_{11} + j, a_{12})}{B(a_{11}, a_{12})} \frac{B(a_{21} + 2h - j, a_{22})}{B(a_{21}, a_{22})}.$$

The default moment prior will be obtained by letting $a_{11} = a_{12} = b_1$ and $a_{21} = a_{22} = b_2$; letting $h = 0$ will then return, as usual, the default prior.

Consider now the intrinsic approach applied to the default moment prior. Since the data consist of two counts, a vector of length two is needed to specify the training sample size. The *intrinsic moment* prior of order h with training sample size $t = (t_1, t_2)$ will be defined as

$$(17) \quad p_1^{IM}(\theta_1, \theta_2|b, h, t) = \sum_{x_1=0}^{t_1} \sum_{x_2=0}^{t_2} p_1^M(\theta_1, \theta_2|a_x^*, h) m_0(x_1, x_2|b_0),$$

where $b = (b_0, b_1, b_2)$, while $(a_x^*)_{11} = b_1 + x_1$, $(a_x^*)_{12} = b_1 + t_1 - x_1$, $(a_x^*)_{21} = b_2 + x_2$, $(a_x^*)_{22} = b_2 + t_2 - x_2$, and

$$(18) \quad m_0(x_1, x_2|b_0) = \binom{t_1}{x_1} \binom{t_2}{x_2} \frac{B(b_0 + x_1 + x_2, b_0 + t_1 + t_2 - x_1 - x_2)}{B(b_0, b_0)};$$

letting $h = 0$ returns the standard intrinsic prior $p_1^I(\theta_1, \theta_2|b, t)$.

The BF against \mathcal{M}_0 using the intrinsic moment prior under \mathcal{M}_1 is given by

$$(19) \quad BF_{10}^{IM}(y_1, y_2|b, h, t) = \sum_{x_1=0}^{t_1} \sum_{x_2=0}^{t_2} BF_{10}^M(y_1, y_2|a_x^*, h) m_0(x_1, x_2|b_0),$$

where $BF_{10}^M(y_1, y_2|a_x^*, h)$ is the BF obtained with the ‘‘posterior’’ $p_1^M(\theta_1, \theta_2|a_x^*, h)$ as parameter prior under \mathcal{M}_1 (and the default parameter prior under \mathcal{M}_0).

Similarly to the Bernoulli case, we can write

$$(20) \quad BF_{10}^M(y_1, y_2|a, h) = \frac{K(a_y^*, h)}{K(a, h)} BF_{10}(y_1, y_2|a),$$

where $(a_y^*)_{11} = a_{11} + y_1$, $(a_y^*)_{12} = a_{12} + n_1 - y_1$, $(a_y^*)_{21} = a_{21} + y_2$, and $(a_y^*)_{22} = a_{22} + n_2 - y_2$. A standard computation then gives

$$m_1(y_1, y_2|a) = \binom{n_1}{y_1} \binom{n_2}{y_2} \frac{B(a_{11} + y_1, a_{12} + n_1 - y_1) B(a_{21} + y_2, a_{22} + n_2 - y_2)}{B(a_{11}, a_{12}) B(a_{21}, a_{22})},$$

and it follows that the Bayes factor against \mathcal{M}_0 obtained with the moment prior under \mathcal{M}_1 (and the default prior under \mathcal{M}_0) can be written as

$$BF_{10}(y_1, y_2|a) = \frac{B(b_0, b_0) B(a_{11} + y_1, a_{12} + n_1 - y_1) B(a_{21} + y_2, a_{22} + n_2 - y_2)}{B(a_{11}, a_{12}) B(a_{21}, a_{22}) B(b_0 + y_1 + y_2, b_0 + n_1 + n_2 - y_1 - y_2)}.$$

Using the above expression in (20) and plugging the latter into (19) provides an explicit expression for $BF_{10}^{IM}(y_1, y_2|b, h, t)$.

4.1 Choice of hyperparameters

The intrinsic moment prior $p_1^{IM}(\theta_1, \theta_2 | b, h, t)$ depends on three hyperparameters. We recommend choosing $b_1 + b_2 = b_0$, so that the same amount of prior information is imposed under \mathcal{M}_1 , on the vector parameter (θ_1, θ_2) , and under \mathcal{M}_0 , on the scalar parameter θ . Specifically, adopting a prior distribution with unit prior information, we let $b_0 = 1/2$, and $b_1 = b_2 = 1/4$, for the balanced case $n_1 = n_2$, while in the non-balanced case b_1 and b_2 will be proportional to n_1 and n_2 . Then, as in the Bernoulli example, we recommend choosing $h = 1$, which is enough to change the asymptotic learning rate of the BF, when the null holds, from sub-linear to super-linear. Finally, concerning the choice of t , we follow the general procedure outlined in the Bernoulli example, with suitable specific modifications to deal with the present case. In particular, we focus on the balanced case to obtain a single optimal value of $t_+ = t_1 + t_2$, which can then be used also in the non-balanced case to specify t_1 and t_2 as (approximately) proportional to n_1 and n_2 .

Clearly $n_1 = n_2 = 1$ represent the minimal sample sizes for the testing problem at hand. In this case, of the four possible data outcomes, two are supportive for \mathcal{M}_0 , namely $(y_1 = 0, y_2 = 0)$ and $(y_1 = 1, y_2 = 1)$, and two are supportive for \mathcal{M}_1 , namely $(y_1 = 0, y_2 = 1)$ and $(y_1 = 1, y_2 = 0)$. We repeat the argument in Subsection 3.1 and take $t_+^* = \operatorname{argmax}_{t_+} TWOE(t_+)$ as the optimal total training sample size, where $TWOE(t_+) = \sum_y WOE_y(t_+)$ and $WOE_y(t_+) = \log BF_{10}^{IM}(y_1, y_2 | b, h, t)$ with $t = (t_+/2, t_+/2)$ and t_+ even.

Figure 3(a) plots $TWOE(t_+)$ for $h = 0, 1, 2$. As in the Bernoulli case, t_+^* is well-defined and when $h = 0$ (standard intrinsic prior) we get $t_+^* = 0$. Hence, in this case, we would recommend a sensitivity analysis in line with that carried out by Casella and Moreno (2009, Table 2). On the other hand, when $h = 1$ we find $t_+^* = 8$, while for $h = 2$ we get $t_+^* = 14$; as in the Bernoulli case, it turns out that starting with a non-local moment prior the intrinsic approach is needed. In the following subsection we highlight some features of the intrinsic moment priors specified by the above values of h and $t_+ = t_+^*$ (including $h = 0$ and $t_+^* = 0$).

4.2 Characteristics of intrinsic moment priors

Figure 4 presents a collection of nine priors for (θ_1, θ_2) under \mathcal{M}_1 , each labelled with its corresponding correlation coefficient r . Although the absolute values of r are of dubious utility in describing these distributions, because of their shape, comparison of the displayed values enables us to highlight the roles played by h and t_+ : as h grows the prior mass is displaced from areas around the line $\theta_1 = \theta_2$ to the corners $(\theta_1 = 0, \theta_2 = 1)$ and $(\theta_1 = 1, \theta_2 = 0)$, thus inducing negative correlation; on the other hand, as t_+ grows the prior mass is pulled back towards either side of the line $\theta_1 = \theta_2$, and positive correlation is induced. The priors in the first row are local, while those in the second and third row are non-local. The three distributions on the main diagonal represent, for the three values of h , our suggested priors based on the criterion for the choice of t_+ described in Subsection 4.1. Notice that $r \simeq 0$ for all three suggested priors, so that the chosen value of t_+ can be seen as ‘‘compensating’’ for h .

Some further insight into the structure of the priors on the main diagonal of Figure 4 can be gleaned by looking at Figure 3(b), which reports their marginal distributions (identical for θ_1 and θ_2). All three densities are symmetric around

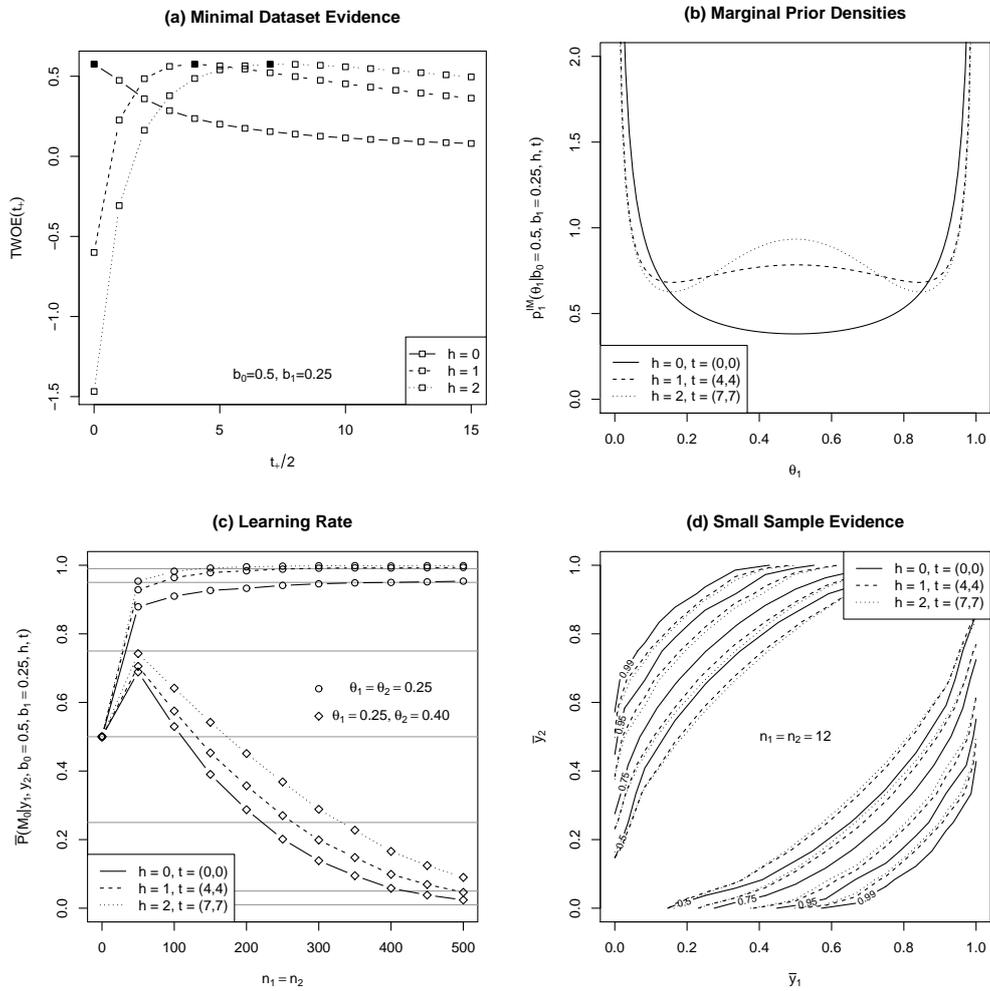


FIG 3. Characteristics of intrinsic moment priors for comparing two proportions. Horizontal gray lines in (c) denote possible decision thresholds at 1%, 5%, 25%, 50%, 75%, 95% and 99% on the posterior probability scale. Contour lines in (d) refer to the posterior probability of the alternative model computed from data $y_1 = n_1 \bar{y}_1$ and $y_2 = n_2 \bar{y}_2$ (letting $b_0 = 1/2$ and $b_1 = 1/4$).

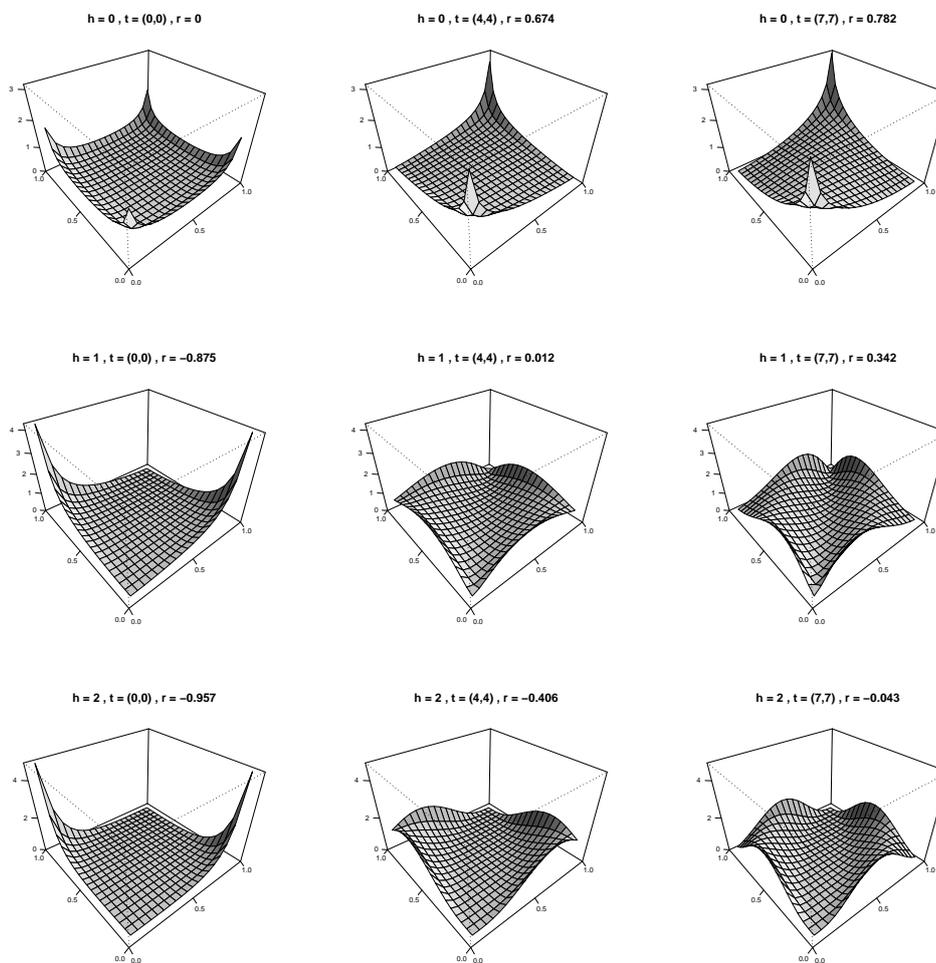


FIG 4. *Intrinsic moment prior densities for comparing two proportions ($b_0 = 1/2, b_1 = 1/4$).*

the value 0.5, but the two intrinsic moment priors with $h > 0$ give more credit to the inner values of the interval $(0, 1)$. For these three priors, Figure 3(c) reports the average posterior probability of the null model computed on 1000 simulated data sets of increasing size generated first letting $\theta_1 = \theta_2 = 0.25$ and then setting $\theta_1 = 0.25$, $\theta_2 = 0.4$ (an instance of the alternative model). Notice that, while in the Bernoulli example we were able to implement an exact computation, in this case we had to resort to a Monte Carlo approximation, because exact computation would have been too demanding (at least for ordinary computational resources).

The learning rate is quite different under the three priors on the main diagonal of Figure 4. Like in the Bernoulli example, when the data are generated under the null model a much quicker correct response is provided by the non-local priors: for sample sizes up to 500 the average posterior probability of \mathcal{M}_0 under the default prior hardly reaches the 95% threshold, whereas under the non-local intrinsic moment priors it easily achieves the 99% threshold by the time 250 observations have been collected. On the other hand, switching from $h = 0$ to $h > 0$, the learning rate under the alternative model is compromised in the short run, but not in the long run.

Figure 3(d) illustrates the small sample behaviour of intrinsic moment priors, by reporting the contour lines in the (\bar{y}_1, \bar{y}_2) -plane of observed frequencies, when $n_1 = n_2 = 12$, for selected thresholds of the posterior probability of \mathcal{M}_1 . There is a clear indication that the displayed thresholds are reached for pairs (\bar{y}_1, \bar{y}_2) closer to the $\bar{y}_1 = \bar{y}_2$ line under the non-local priors than under the default prior. Similarly to the Bernoulli example, this is due to the steeper gradient of the evidence surface as the data move away from the null supporting values.

5. VARIABLE SELECTION IN LOGISTIC REGRESSION MODELS

We now develop the intrinsic moment procedure when the models under comparison are logistic regression models. This demonstrates that the general procedure can be applied to a flexible and general class of discrete data models.

Suppose we observe N independent binomial observations, $y = (y_1, \dots, y_N)$, where

$$y_i | \theta_i \sim \text{Bin}(n_i, \theta_i); \quad i = 1, \dots, N.$$

The binomial probabilities $\theta = (\theta_1, \dots, \theta_N)$ are assumed to depend on the values of k explanatory variables z_{ij} , $i = 1, \dots, N$, $j = 1, \dots, k$, through linear predictors $\eta = (\eta_1, \dots, \eta_N)$, where

$$\eta_i = \log \frac{\theta_i}{1 - \theta_i} = \beta_0 + \sum_{j=1}^k z_{ij} \beta_j; \quad i = 1, \dots, N.$$

Hence the likelihood is $f(y | \beta) = \left\{ \prod_{i=1}^N \binom{n_i}{y_i} \right\} L(\beta | y, n)$, where

$$L(\beta | y, n) = \exp \left\{ y_i \left(\beta_0 + \sum_{j=1}^k z_{ij} \beta_j \right) - n_i \log \left(1 + \exp \left[\beta_0 + \sum_{j=1}^k z_{ij} \beta_j \right] \right) \right\},$$

$\beta = (\beta_0, \beta_1, \dots, \beta_k)$ and $n = (n_1, \dots, n_N)$. We refer to this model as the full model. Further models under consideration for variable selection correspond to an exclusion of some explanatory variables, that is, to setting some $\beta_j = 0$ ($j \neq 0$).

In the development below, we present the prior for the full model with k explanatory variables, but the prior for any other model takes an identical form with a regression parameter of correspondingly lower dimensionality.

For convenience, and consistency with our earlier developments in the context of two binomial models, we adopt a conjugate local prior (Bedrick, Christensen and Johnson, 1996) given by $p^C(\beta | u, w) \propto L(\beta | u, w)$, where $u = (u_1, \dots, u_N)$ and $w = (w_1, \dots, w_N)$ are hyperparameters corresponding to $y = (y_1, \dots, y_N)$ and $n = (n_1, \dots, n_N)$, respectively, in the likelihood. Letting $u_+ = \sum_{i=1}^N u_i$, we choose as the default prior specification

$$(21) \quad w_i = w_+ \frac{n_i}{\sum_i n_i}, \quad u_i = \frac{w_i}{2},$$

where w_+ represents a prior sample size. The condition $u_i = w_i/2$ ensures that the mode of the prior is at $\beta = 0$. To see why, recall that the prior, as a function of β , is proportional to the likelihood. Now, if $y_i = n_i/2$, then the MLE of each θ_i , unconstrained by the model, is exactly $1/2$, and therefore the MLE of each η_i is zero. The value $\eta = 0$ is attained within any logistic regression model by $\beta = 0$ and hence this value must also maximize the model constrained likelihood, which corresponds to the prior density. As a default choice, corresponding to unit prior information, we take $w_+ = 1$. For the comparison of two proportions θ_1 and θ_2 ($N = 2$) in the balanced case $n_1 = n_2$, this formulation leads to identical default local priors $\theta_i \sim \text{Beta}(1/4, 1/4)$ with θ_1 and θ_2 independent.

In order to construct the moment prior, we need to specify a function $g_h(\beta)$. We choose

$$(22) \quad g_h(\beta) = \prod_{j=1}^k \beta_j^{2h},$$

which vanishes if at least one $\beta_j = 0$ ($j \neq 0$) implying that we separate the full model from every model nested within it having one fewer explanatory variable. In the context of variable selection for Gaussian distributions, this choice of $g_h(\beta)$ has been used by Consonni and La Rocca (2011) and also by Johnson and Rossell (2012), who named the resulting non-local prior a *product moment* prior. Our main result in the Appendix (Theorem A.2), though stated for i.i.d. observations, confirms that this is a sensible choice for variable selection, resulting in an effective separation of models. With this choice we obtain

$$p^M(\beta | u, v, h) \propto p^C(\beta | u, w) \prod_{j=1}^k \beta_j^{2h}.$$

At this stage, to specify the intrinsic moment prior under any given model \mathcal{M} , we need a reference model \mathcal{M}_0 , which we take as the null model having no explanatory variable ($k = 0$) because it is nested in every other model. In this construction, the priors used in any pairwise model comparison depend only on the (common) null model. This strategy is called *encompassing from below* and provides a coherent model comparison procedure; see Liang et al. (2008). Under \mathcal{M}_0 we assume a default prior for the intercept β_0 given by

$$p(\beta_0) \propto \exp\{\beta_0 u_+ - w_+ \log(1 + \exp[\beta_0])\},$$

where $u_+ = \sum_{i=1}^N u_i$, which corresponds to a $Beta(u_+, w_+ - u_+) = Beta(1/2, 1/2)$ distribution, because of the assumed value $w_+ = 1$, for the common success probability implied by \mathcal{M}_0 in the comparison of two proportions.

The final step in the construction of the intrinsic moment prior requires the specification of training samples, which also involves covariates when dealing with regression models. Methods for choosing covariates for training data have been discussed for Gaussian regression models by Girón et al. (2006). We assume that the covariate patterns in the training data are a subset of those appearing in the observed data. Following formula (2), we now construct the intrinsic moment prior for the parameter of a logistic regression model. Let

$$p^M(\beta | x + u, t + w, h) \propto \left\{ \prod_{j=1}^k \beta_j^{2h} \right\} L(\beta | x + u, t + w)$$

be the posterior moment prior based on the training sample $x = (x_1, \dots, x_N)$, with training sample sizes given by $t = (t_1, \dots, t_N)$, where some of the t_i may be zero. Since x is drawn from $m_0(x)$, the marginal joint distribution under \mathcal{M}_0 , the intrinsic moment prior is thus given by

$$p^{IM}(\beta | h, t) = \sum_x m_0(x) \frac{\{\prod_{j=1}^k \beta_j^{2h}\} L(\beta | x + u, t + w)}{Q(x + u, t + w, h)},$$

where $Q(z, s, h) = \int_{\mathfrak{R}^{k+1}} \{\prod_{j=1}^k \beta_j^{2h}\} L(\beta | z, s) d\beta$. The existence of $Q(z, s, h)$ follows from the theorem in Forster (2010, Section 6) stating that a necessary and sufficient condition for a log-concave function over \mathcal{R}^d to have a finite integral is that it achieves its maximum in the interior of the parameter space. Here, we need to adapt this result slightly. Firstly, we notice that in each (open) orthant, the integrand is log-concave, because both its constituent components are: $\prod_{j=1}^k \beta_j^{2h}$ by straightforward calculus, and $L(\beta | x + u, t + w)$ by log-concavity of the likelihood for a binomial logistic regression model. For our default choices of u and w (or any alternative choice with $u > 0$ and $0 < u < w$) $L(\beta | x + u, t + w)$ has a unique finite maximum, provided that the model is identified (which we will assume). Hence $L(\beta | x + u, t + w)$ tends to zero, as $\|\beta\| \rightarrow \infty$, in any direction, and so does $\{\prod_{j=1}^k \beta_j^{2h}\} L(\beta | x + u, t + w)$, due to the dominance of $L(\beta | x + u, t + w)$ for large $\|\beta\|$. Hence, we have the conditions to apply the result of Forster (2010) in each orthant to guarantee a finite integral.

Now we require to compute the marginal likelihood induced by $p^{IM}(\beta | h, t)$. This is given by

$$\begin{aligned} m^{IM}(y | h, t) &= \left\{ \prod_{i=1}^N \binom{n_i}{y_i} \right\} \int_{\mathfrak{R}^{k+1}} L(\beta | y, n) p^{IM}(\beta | h, t) d\beta \\ (23) \quad &= \left\{ \prod_{i=1}^N \binom{n_i}{y_i} \right\} \sum_x m_0(x) \frac{Q(x + u + y, t + w + n, h)}{Q(x + u, t + w, h)}. \end{aligned}$$

In practice, we need an efficient method to compute $Q(z + x, s + t, h)$ for $(z, s) =$

(u, w) and $(z, s) = (u + y, w + n)$. Since

$$\begin{aligned} Q(z + x, s + t, h) &= \int_{\mathfrak{R}^{k+1}} \frac{L(\beta | z, s)L(\beta | x, t)}{Q(z, s, 0)} Q(z, s, 0) \left\{ \prod_{j=1}^k \beta_j^{2h} \right\} d\beta \\ &= Q(z, s, 0) \mathbb{E}^{p^C(\beta | z, s)} \left\{ \left(\prod_{j=1}^k \beta_j^{2h} \right) L(\beta | x, t) \right\}, \end{aligned}$$

one can simulate from the conjugate local prior $p^C(\beta | z, s)$ using MCMC methods and obtain $m^{IM}(y | h, t)$ as a mixture, with respect to $m_0(x)$, of ratios of expectations; the normalizing constants $Q(z, s, 0)$ for $h = 0$ will be computed once and for all, for a given data set, again using MCMC methods.

6. APPLICATIONS

In this section we apply our methodology to two problems. The first application concerns a set of randomized trials, and uses results presented in Section 4; the second application performs model selection within a logistic regression framework to analyze the relationship between the probability of patients' survival and two binary covariates, and makes use of results presented in Section 5.

6.1 Randomized trials

We analyze data from 41 randomized trials of a new surgical treatment for stomach ulcers. For each trial the number of occurrences and nonoccurrences under Treatment (the new surgery, *group 1*) and Control (an older surgery, *group 2*) are reported; see Efron (1996, Table 1). Occurrence here refers to an adverse event: recurrent bleeding. Efron (1996) analyzed these data with the aim of performing a meta-analysis, using empirical Bayes methods. On the other hand, our objective is to establish whether the probability of occurrence is the same under Treatment and Control in each individual table; for a similar analysis see Casella and Moreno (2009). We base our analysis on the intrinsic moment priors of Section 4, letting $b_0 = 1/2$ and comparing the results given by different choices of h and t . Specifically, we perform a sensitivity analysis with respect to the actual choice of t , and a cross-validation study of the predictive performance achieved by different choices of h .

6.1.1 Sensitivity analysis Recall that a crucial hyperparameter is represented by the overall training sample size t_+ , which is then further split into the two groups, $t_+ = t_1 + t_2$. In Subsection 4.2, on the basis of our study of the characteristics of intrinsic moment priors for the comparison of two proportions, we suggested a sensitivity analysis with $t_+ > t_+^*(h)$. Accordingly, we here let t_+ vary from $t_+^*(h)$ to $t_+^{**}(h) = t_+^*(h + 1)$, where $h = 0$ (standard local prior) or $h = 1$ (recommended non-local prior); recall that $t_+^*(0) = 0$, $t_+^*(1) = 8$ and $t_+^*(2) = 14$. We choose t_1 and t_2 approximately proportional to the trial sample sizes for Treatment and Control, n_1 and n_2 , and b_1 and b_2 exactly proportional to these quantities, with $b_1 + b_2 = b_0 = 1/2$ (unit prior information). For all the above pairs (h, t) , and all 41 tables in the data set, we evaluate the posterior probability of the null model.

We report our findings in Figure 5(a), where the tables are arranged (for a better appreciation of our results) from left to right in increasing order of $|\frac{y_1}{n_1} - \frac{y_2}{n_2}|$

(absolute difference in observed fractions): this explains the mostly declining pattern of the posterior probabilities of the null model. The range of these probabilities is depicted as a vertical segment, separately for the standard intrinsic and the intrinsic moment prior, and the values for $t = t^*$ and $t = t^{**}$ are marked with circles and triangles, respectively, so that in most cases (thanks to a monotonic behavior) we can see an arrow describing the overall change in probability. One can identify three sets of tables: left-hand (up to table 38), center (tables from 20 to 7) and right-hand (remaining tables). Some specific comments follow below.

Consider first the left-hand tables. Except for table 41 under the local prior (and possibly table 18) the posterior probability of \mathcal{M}_0 ranges well above the value 0.5, which can be regarded as a conventional decision threshold for model choice under a $\{0,1\}$ -loss function. The non-local intrinsic moment prior (black triangle) produces values for the posterior probability of \mathcal{M}_0 higher than under the standard intrinsic prior (white triangle): this is only to be expected, because of the non-local *versus* local nature of these priors. The effect is dramatic for table 41, which is characterized by counting no occurrences at all. All arrows point downwards: this is the effect of the intrinsic procedure; when the data support the null model, the action of pulling the prior towards the null subspace makes the alternative more competitive and takes evidence away from \mathcal{M}_0 . For this first group of tables, a robust conclusion can be reached in favor of the equality of proportions between the two groups. Next consider the tables in the center. Four of these tables (20, 39, 15, 7) exhibit a posterior probability of the null hovering over the 0.5 threshold, so that no robust conclusion can be drawn; four of them (32, 26, 16, 33) give a robust conclusion in favor of the null, while two of them (34, 5) give a robust conclusion against the null. Leaving aside these last two tables, which are characterized by zero occurrences in one of the two groups and are more similar in behaviour to the right-hand tables, all arrows point downwards, indicating that here too the intrinsic procedure is working in favor of the alternative. Notice that now the local priors give more credit to the null than the non-local priors, indicating that the steeper evidence gradient of the latter gets into play. Finally, the pattern of the right-hand tables indicates a low support for the null, with the possible exception of tables 1 and 12. Ranges become shorter, and on some occasions negligible, especially for the non-local priors. Some arrows point upwards: this is the action of the intrinsic procedure in favor of \mathcal{M}_0 , because the data do not support the null model; the two Bayesian models are becoming equivalent. For the tables in this last group, a robust conclusion against the null can be drawn.

6.1.2 Cross-validation study We now compare the predictive performance of the intrinsic moment priors with $h = 0$, $h = 1$ and $h = 2$, taking for granted that t_+ should be equal to t_+^* (for any given value of h) and t_1 and t_2 should be (approximately) proportional to n_1 and n_2 . To this aim, we assign a logarithmic score to each probability forecast p , say, of an event E : the score is $\log(p)$, if E occurs, and $\log(1 - p)$, if \bar{E} occurs; this is a proper scoring rule (Bernardo and Smith, 1994, Sect. 2.7.2). Notice that each score is negative, the maximum value it can achieve is zero, and higher scores indicate a better prediction. Suppose we want to predict the outcome for a patient who is an occurrence in group 1. We exclude this patient from the data set and compute the probability for an occurrence of such a patient, $\hat{\theta}_1^{(1)}$, as the Bayesian model average of the posterior

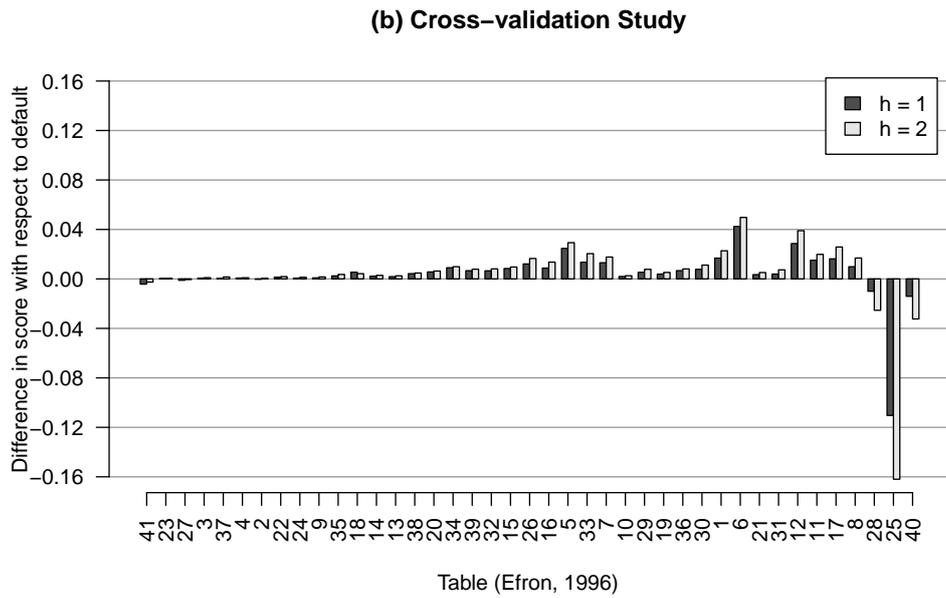
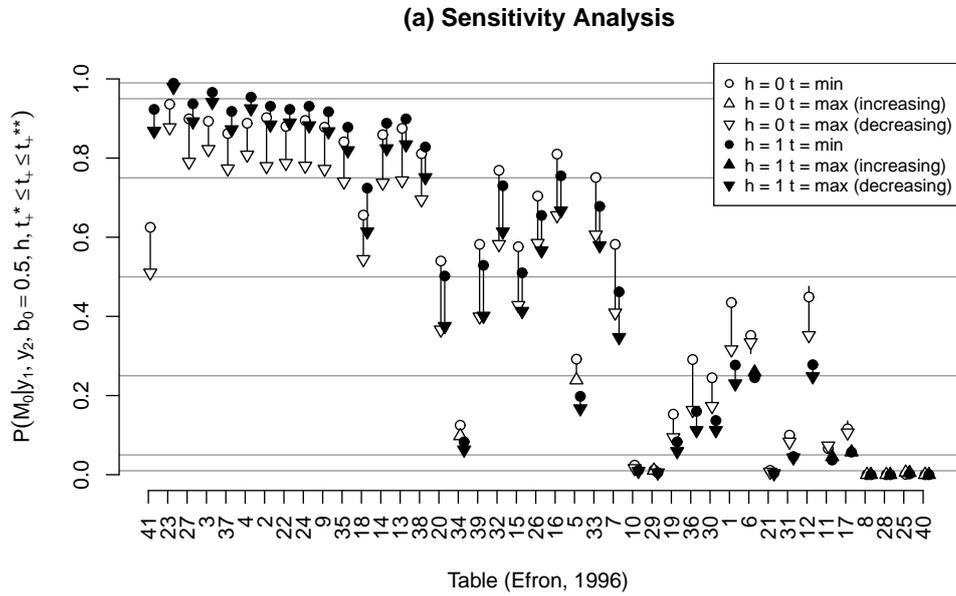


FIG 5. Results of sensitivity analysis and cross-validation study: each number on the horizontal axis identifies a table.

means of θ_1 under \mathcal{M}_1 and θ under \mathcal{M}_0 based on counts $(y_1 - 1, n_1 - y_1, y_2, n_2 - y_2)$; similarly, to predict the outcome for a patient who is an occurrence in group 2, we compute her probability of occurrence, $\hat{\theta}_2^{(1)}$, upon interchanging subscript 1 and 2 above. On the other hand, to predict the outcome for a patient who is a nonoccurrence in group 1, we compute the corresponding probability of an occurrence, $\hat{\theta}_1^{(0)}$, as the Bayesian model average of the posterior means of θ_1 under \mathcal{M}_1 and θ under \mathcal{M}_0 based on counts $(y_1, n_1 - y_1 - 1, y_2, n_2 - y_2)$; as before, the computation of $\hat{\theta}_2^{(0)}$, for a patient who is a nonoccurrence in group 2, requires interchanging subscript 1 and 2. In the spirit of cross-validation, we repeat the analysis for each patient and compute the overall mean score

$$S = \frac{y_1 \log \hat{\theta}_1^{(1)} + (n_1 - y_1) \log(1 - \hat{\theta}_1^{(0)}) + y_2 \log \hat{\theta}_2^{(1)} + (n_2 - y_2) \log(1 - \hat{\theta}_2^{(0)})}{n_1 + n_2}.$$

Now let S_h be the score associated with the intrinsic moment prior of order h , $h = 0, 1, 2$. Of particular interest are the differences $S_1 - S_0$ and $S_2 - S_0$. A positive value for $S_1 - S_0$, say, means that the prior with $h = 1$ produces on average a better forecasting system than the standard intrinsic prior ($h = 0$); notice that the latter coincides with the default prior because $t_+^* = 0$. One can use a first order expansion of the logarithmic score to gauge the difference more concretely: a positive difference $S_1 - S_0 = d > 0$ means that the prior with $h = 1$ generates ‘‘correctly-oriented probability forecasts’’ (higher values for occurrences and lower values for nonoccurrences) which are, on average, $d \times 100\%$ better than those produced by the standard intrinsic prior. Here the average is taken over the combination of event outcomes (occurrence/nonoccurrence) and groups (Treatment/Control) with weights given by the observed sample frequencies. Since $d > 0$ is an average of score differences over the four blocks of events, there is no guarantee of a uniform improvement in prediction across all of them.

Figure 5(b) reports the results of our cross-validation study with the tables again arranged from left to right in increasing order of absolute difference in observed fractions. Essentially for all tables, but with the notable exception of the last three, the non-local intrinsic moment priors perform better than the standard intrinsic prior, with differences in score ranging from -0.42% to 4.2% (median improvement 0.54%) when $h = 1$ and from -0.26% to 5.0% (median improvement 0.68%) when $h = 2$. On the other hand, for the last three tables, which are clearly against the null, the performance of non-local priors is much worse: this happens because the intrinsic moment priors produce a greater degree of posterior shrinkage towards the null within the alternative model. Differences in score range from -1.0% down to -11% , when $h = 1$, and from -2.5% down to -16% , when $h = 2$. Notice that the intrinsic moment prior predicts better with $h = 2$ than with $h = 1$ when the difference in score is positive, but the reverse occurs for negative differences in score; in the latter case the performance can be appreciably worse. On grounds of prudence, these results seem to reinforce our recommendation in favor of the choice $h = 1$.

6.2 Logistic regression models for survival data

In Table 1 we consider a data set previously examined in Dellaportas, Forster and Ntzoufras (2002); see also references therein for further analyses of the same problem. Our aim is to investigate the relationship between the probability of

TABLE 1
Survival data

	Antitoxin	Death	Survival
More Severe	Yes	15	6
	No	22	4
Less Severe	Yes	5	15
	No	7	5

Survival on the one hand, and two binary covariates: Severity of condition and Antitoxin medication.

The full model is given by

$$y_{jl} | \theta_{jl} \stackrel{ind}{\sim} \text{Bin}(n_{jl}, \theta_{jl}),$$

$$\log\left(\frac{\theta_{jl}}{1 - \theta_{jl}}\right) = \alpha + \beta_j + \gamma_l + \delta_{jl},$$

$j, l = 1, 2$, where y_{jl} , n_{jl} and θ_{jl} are the number of survivals, the total number of patients and the probability of survival under level j of Severity and level l of Antitoxin medication; α , β_j , γ_l and δ_{jl} are the model parameters corresponding to the intercept, Severity effect, Antitoxin effect, and interaction effect of Severity and Antitoxin. The number of free parameters is actually four: intercept, two main effects and one interaction.

We are interested in five distinct logistic regression models: the intercept-only model, two models with a single main effect each (plus intercept), one model with two additive main effects (plus intercept), and the full model. We wish to compare them through their posterior probabilities based on our intrinsic moment priors. Our results are summarized in Table 2, where we report posterior model probabilities with an accuracy (standard error) of approximately 1%.

Computations were performed using the methodology presented in Section 5. In particular, for the choice of prior hyperparameters, we used formula (21) with $w_+ = 1$. A uniform prior on the model space was assumed. For each model, a random walk Metropolis-Hastings sampler was implemented through the function `metrop()` of the R package `mcmc` (Geyer, 2010). Prior and posterior normalizing constants with $h = 0$ were computed, once and for all, using the method by Chib and Jeliazkov (2001) on chains of length 40000 after thinning by a factor 20; the proposal distributions were tuned so as to obtain acceptance rates between 24% and 28%. Different chains with the same features were used to compute the ratios of posterior expectations needed to calculate the intrinsic moment marginal likelihood (23) as a mixture with respect to $m_0(x)$. Since the mixing step proved to be computationally demanding, we used to C (within R).

Differently from the case of the comparison of two proportions, for general logistic regression models there seems to be no simple method to determine t_+^* once and for all, because the explanatory variables are different in each application. Moreover, extending the methodology of total weight of evidence presented in Subsection 3.1 to the case of more than two models appears to be non-trivial. In the present application we found it natural to let $n_{j,l} \equiv 1$ for minimal data, and contented ourselves with computing the total weight of evidence for the full model against the intercept-only one, focussing on the two models farthest from each other. We used this information to guide our choice of t_+ in the context of

TABLE 2

Posterior probabilities of five logistic regression models for the survival data in Table 1, using intrinsic moment priors with total weight of evidence on corresponding minimal data in the last column. Each model is described through the main effect(s) it includes beside intercept.

h	t_+	Intercept-Only	Severity	Antitoxin	Sever + Antitox	Full model	TWOE
0	0	0.01	0.61	0.01	0.35	0.02	7
	4	0.01	0.56	0.01	0.40	0.01	4
	8	0.00	0.44	0.01	0.51	0.03	3
	12	0.00	0.35	0.01	0.54	0.10	2
	16	0.00	0.33	0.01	0.54	0.12	2
	20	0.00	0.29	0.01	0.53	0.17	2
	24	0.00	0.26	0.01	0.52	0.21	2
1	0	0.22	0.77	0.01	0.00	0.00	2
	4	0.03	0.86	0.01	0.10	0.00	7
	8	0.01	0.85	0.01	0.13	0.00	6
	12	0.00	0.67	0.01	0.31	0.00	6
	16	0.00	0.62	0.01	0.36	0.00	6
	20	0.00	0.52	0.01	0.46	0.00	6
	24	0.00	0.45	0.01	0.53	0.00	5
2	0	0.95	0.05	0.00	0.00	0.00	-2
	4	0.13	0.86	0.01	0.01	0.00	21
	8	0.03	0.95	0.01	0.00	0.00	25
	12	0.01	0.93	0.01	0.05	0.00	26
	16	0.01	0.89	0.01	0.09	0.00	27
	20	0.00	0.80	0.01	0.19	0.00	26
	24	0.00	0.70	0.01	0.29	0.00	25

a sensitivity analysis across a grid of values for the hyperparameters $h = 0, 1, 2$ and $t_+ = 0, 4, 8, 12, 16, 20, 24$; the actual values of t_{jl} were obtained, by rounding them, as approximately proportional to n_{jl} . In general, the choice of t could depend on the model, which could help comparing models of very different dimension, but in the present case we avoided this additional complexity.

The values of the total weight of evidence in Table 2 suggest that we should take $t_+ = 0$ when $h = 0$, $t_+ = 4$ when $h = 1$, and $t_+ = 16$ when $h = 2$. However, the last column of Table 2 is not stable across different MCMC runs, and it should be considered as merely indicative. This is not surprising, because the total weight of evidence was quite flat around its maximum in both Figure 2(b) and Figure 3(a); it is a problem that cannot be solved by a feasible increase in chain length. The clear message appears to be that $t_+ = 12$ is too much when $h = 0$, $t_+ = 0$ is not a good choice when $h = 1$, and $t_+ = 4$ is not enough when $h = 2$. Notice that the first value would give some credit to the full model, while the last two values would attribute a sizeable posterior probability to the intercept-only model. Then, if a recommended value $t_+^*(h)$ has to be singled out for each value of h , the choice $t_+^*(0) = 0$, $t_+^*(1) = 8$ and $t_+^*(2) = 16$ achieves a better scaling with respect to h , and it is in line with the values found for the comparison of two proportions. Here too the intrinsic step appears to be necessary for non-local priors only.

Bearing in mind that the intercept term is present in each model, the posterior model probabilities reported in Table 2 suggest that the two models ‘‘Severity’’ and ‘‘Severity+Antitoxin’’ account for at least 90% of the probability mass in all reasonable scenarios. Specifically, model ‘‘Severity’’ is a clear winner under the

non-local priors, except for $h = 1$ and the highest values of t_+ , which are far from $t_+^*(1)$. The situation is more mixed under the local priors: the leadership of “Severity” is not equally clear, and it fades away as t_+ increases; these results are in line with those obtained by Dellaportas, Forster and Ntzoufras (2002) using several MCMC schemes all based on local normal parameter priors. While local and non-local priors broadly agree on the two leading models, they diverge on the allocation of probability mass between them: for values of t_+ close to the recommended ones non-local priors more sharply select the parsimonious model “Severity”, dropping its more complex competitor “Severity+Antitoxin”.

7. DISCUSSION

In this paper we have presented a general approach to objective Bayesian testing for nested hypotheses in discrete data models. The only required input is a default (proper) parameter prior under each of the entertained models. Next, a default non-local prior is derived, and finally a procedure based on the intrinsic methodology is applied. The fundamental tool in our approach is represented by a particular class of non-local priors, which we name intrinsic moment priors. These distributions combine the virtues of non-local priors and intrinsic priors to obtain balanced objective tests, whose learning rate is improved (strongly accelerated when the smaller model holds) relative to current local prior methods, while their small sample evidence is broadly comparable with that afforded by modern objective methods, including those based on intrinsic priors.

An important feature of intrinsic moment priors is represented by the training sample size. We handle the choice of this hyperparameter in a novel way, and quite differently from current intrinsic approaches, using the notion of total weight of evidence. This criterion looks promising, at least for finitely discrete data models, but it cannot be naively extended to the countably infinite or continuous case, because we cannot weight data values uniformly; a suitable weighting data measure should be devised, whose choice however remains an open issue. Whether or not an optimal value for the training sample size can be found, one can always carry out a sensitivity analysis; an exercise we typically recommend to assess robustness of conclusions with respect to this hyperparameter.

Our approach for the construction of prior distributions is based on a comparison of two nested models. When several models are entertained, we select the null model as a natural baseline, because it is nested within any other model, similarly to methods based on intrinsic priors (Girón et al., 2006) or on mixtures of g -priors (Liang et al., 2008). This choice of course can be modified, if an alternative minimal model is available. Clearly, the baseline model acquires a special status in this approach. Assignments of parameter priors for pairwise comparison of models which are symmetric in nature, because they do not require a baseline model, are developed in Cano, Salmerón and Robert (2008).

While our analysis was solely based on proper priors, we emphasize that moment priors can also be improper. The subsequent analysis can then proceed through an intrinsic step, as in this paper, or through other methods currently available to deal with improper priors for model comparison, such as expected posterior priors, or fractional Bayes factors (O’Hagan, 1995); for an application of the latter methodology see Consonni and La Rocca (2011) and Altomare, Consonni and La Rocca (2013).

REFERENCES

- ALTOMARE, D., CONSONNI, G. and LA ROCCA, L. (2013). Objective Bayesian Search of Gaussian Directed Acyclic Graphical Models for Ordered Variables with Non-Local Priors. *Biometrics*. Early view. DOI: 10.1111/biom.12018.
- BEDRICK, E. J., CHRISTENSEN, R. and JOHNSON, W. (1996). A New Perspective on Priors for Generalized Linear Models. *Journal of the American Statistical Association* **91** 1450–1460.
- 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.
- BERNARDO, J. M. and SMITH, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons.
- CANO, J. A., SALMERÓN, D. and ROBERT, C. P. (2008). Integral Equation Solutions as Prior Distributions for Bayesian Model Selection. *TEST* **17** 493–504.
- CASELLA, G. and MORENO, E. (2005). Intrinsic Meta-analysis of Contingency Tables. *Statistics in Medicine* **24** 583–604.
- CASELLA, G. and MORENO, E. (2006). Objective Bayesian Variable Selection. *Journal of the American Statistical Association* **101** 157–167.
- CASELLA, G. and MORENO, E. (2009). Assessing Robustness of Intrinsic Tests of Independence in Two-Way Contingency Tables. *Journal of the American Statistical Association* **104** 1261–1271.
- CASELLA, G., GIRÓN, F. J., MARTÍNEZ, M. L. and MORENO, E. (2009). Consistency of Bayesian Procedures for Variable Selection. *Annals of Statistics* **37** 1207–1228.
- CHIB, S. and JELIAZKOV, I. (2001). Marginal Likelihood From the Metropolis-Hastings Output. *Journal of the American Statistical Association* **96** 270–281.
- COHEN, J. (1992). A Power Primer. *Psychological Bulletin* **112** 155–159.
- CONSONNI, G. and LA ROCCA, L. (2008). Tests Based on Intrinsic Priors for the Equality of Two Correlated Proportions. *Journal of the American Statistical Association* **103** 1260–1269.
- CONSONNI, G. and LA ROCCA, L. (2011). On Moment Priors for Bayesian Model Choice with Applications to Directed Acyclic Graphs. In *Bayesian Statistics 9 – Proceedings of the Ninth Valencia International Meeting* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 119–144. Oxford University Press With discussion.
- CONSONNI, G., MORENO, E. and VENTURINI, S. (2011). Testing Hardy-Weinberg Equilibrium: An Objective Bayesian Analysis. *Statistics in Medicine* **30** 62–74.
- DAWID, A. P. (2011). Posterior Model Probabilities. In *Philosophy of Statistics* (P. S. Bandyopadhyay and M. Forster, eds.) 607–630. Elsevier.
- DELLAPORTAS, P., FORSTER, J. J. and NTZOUFRAS, I. (2002). On Bayesian Model and Variable Selection Using MCMC. *Statistics and Computing* **12** 27–36.
- EFRON, B. (1996). Empirical Bayes Methods for Combining Likelihoods. *Journal of the American Statistical Association* **91** 538–550. With discussion: 551–565.
- FERGUSON, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall.
- FORSTER, J. J. (2010). Bayesian Inference for Poisson and Multinomial Log-Linear Models. *Stat. Methodol.* **7** 210–224.
- GEYER, C. J. (2010). mcmc: Markov Chain Monte Carlo R package version 0.8.
- GIRÓN, F. J., MORENO, E. and CASELLA, G. (2007). Objective Bayesian Analysis of Multiple Changepoints for Linear Models. In *Bayesian Statistics 8* (J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) 227–252. Oxford University Press With discussion.
- GIRÓN, F. J., MARTÍNEZ, M. L., MORENO, E. and TORRES, F. (2006). Objective Testing Procedures in Linear Models: Calibration of the p -values. *Scand. J. Statist.* **33** 765–784.
- GOUTIS, C. and ROBERT, C. P. (1998). Model Choice in Generalized Linear Models: a Bayesian Approach via Kullback-Leibler Projections. *Biometrika* **85** 29–37.
- JAYNES, E. T. (1979). Review of “Inference, method and decision: Towards a Bayesian philosophy of science”. *Journal of the American Statistical Association* **74** 740–741.
- JEFFERYS, W. H. and BERGER, J. O. (1992). Ockham’s Razor and Bayesian Analysis. *American Scientist* **80** 64–72.
- JEFFREYS, H. (1961). *Theory of Probability*, Third ed. Oxford University Press Corrected impression, 1966.
- JOHNSON, V. E. and ROSSELL, D. (2010). On the Use of Non-Local Prior Densities in Bayesian Hypothesis Tests. *Journal of the Royal Statistical Society, Series B: Methodological* **72** 143–170.

- JOHNSON, V. E. and ROSSELL, D. (2012). Bayesian Model Selection in High-dimensional Settings. *Journal of the American Statistical Association* **107** 649-660.
- KASS, R. E. and RAFTERY, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association* **90** 773-795.
- LEON-NOVELO, L., MORENO, E. and CASELLA, G. (2012). Objective Bayes Model Selection in Probit Models. *Statistics in Medicine* **31** 353-365.
- LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association* **103** 410-423.
- MORENO, E. (1997). Bayes Factors for Intrinsic and Fractional Priors in Nested Models. Bayesian Robustness. In *L₁-Statistical Procedures and Related Topics* (Y. DODGE, ed.) 257-270. Institute of Mathematical Statistics.
- MORENO, E., BERTOLINO, F. and RACUGNO, W. (1998). An Intrinsic Limiting Procedure for Model Selection and Hypotheses Testing. *Journal of the American Statistical Association* **93** 1451-1460.
- MORENO, E., CASELLA, G. and GARCIA-FERRER, A. (2005). An Objective Bayesian Analysis of the Change Point Problem. *Stochastic Environmental Research and Risk Assessment* **19** 191-204.
- MORENO, E., GIRÓN, F. J. and CASELLA, G. (2010). Consistency of Objective Bayes Factors as the Model Dimension Grows. *Annals of Statistics* **38** 1937-1952.
- MORRIS, C. N. (1987). Comments on "Testing a Point Null Hypothesis: The Irreconcilability of *P* Values and Evidence". *Journal of the American Statistical Association* **82** 131-133.
- NEAL, R. M. (2001). Transferring Prior Information Between Models Using Imaginary Data Technical Report No. 0108, Department of Statistics, University of Toronto.
- O'HAGAN, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Methodological)* **57** 99-138.
- O'HAGAN, A. and FORSTER, J. (2004). *Kendall's Advanced Theory of Statistics, Vol. 2b: Bayesian Inference*, Second ed. Arnold.
- PÉREZ, J. M. and BERGER, J. O. (2002). Expected Posterior Prior Distributions for Model Selection. *Biometrika* **89** 491-512.
- ROBERT, C. P. (2001). *The Bayesian Choice: from Decision-theoretic Foundations to Computational Implementation*, Second ed. Springer-Verlag.
- ROBERT, C. P. and CASELLA, G. (1994). Distance Penalized Losses for Testing and Confidence Set Evaluation. *TEST* **3** 163-182.
- ROBERT, C. P., CHOPIN, N. and ROUSSEAU, J. (2009). Rejoinder: Harold Jeffreys's Theory of Probability Revisited. *Statistical Science* **24** 191-194.
- RUBIN, H. (1987). A Weak System of Axioms for Rational Behavior and the Nonseparability of Utility from Prior. *Statistics & Decisions* **5** 47-58.
- SMITH, A. F. M. and SPIEGELHALTER, D. J. (1980). Bayes Factors and Choice Criteria for Linear Models. *Journal of the Royal Statistical Society, Series B: Methodological* **42** 213-220.

APPENDIX A: BAYES FACTOR ASYMPTOTICS

We study the asymptotic learning rate of BFs for comparing nested models, allowing for non-local parameter priors, under fairly general assumptions. It will be understood that the data be discrete, but this assumption will not be crucial to our results. We start with the notion of a regular, possibly mis-specified, model.

DEFINITION A.1. *A non-singleton statistical model $\mathcal{M} = \{f^n(\cdot|\xi), \xi \in \Xi\}$ for a sequence of data $y^{(n)} = (y_1, \dots, y_n)$ taking values in \mathcal{Y}^n is regular with respect to the strictly positive sampling density $q^n(\cdot)$ if the following assumptions hold:*

1. Ξ is an open subset of \mathbb{R}^d ;
2. $f(\cdot|\xi)$ is strictly positive, for all $\xi \in \Xi$;
3. the Kullback-Leibler projection of $q(\cdot)$ on \mathcal{M} is well-defined, that is, there exists a unique $\xi^* \in \Xi$ such that $K_q(\xi^*) = \inf_{\xi \in \Xi} K_q(\xi)$, where $K_q(\xi) = E_q \log\{q(y_1)/f(y_1|\xi)\}$ is the Kullback-Leibler divergence from $q(\cdot)$ to $f(\cdot|\xi)$;

4. $\mathbb{V}_q \log \{q(y_1)/f(y_1|\xi^*)\} < \infty$;
5. the unit log-likelihood $\ell(y_1|\cdot) = \log f(y_1|\cdot)$ is twice continuously differentiable on Ξ with gradient $s(y_1|\cdot)$ and Hessian matrix $H(y_1|\cdot)$, for all $y_1 \in \mathcal{Y}$;
6. $\mathbb{E}_q|\ell(y_1|\xi^*)| < \infty$ and $\mathbb{E}_q\|s(y_1|\xi^*)\|_2^2 < \infty$;
7. there exist a spheric neighbourhood B of ξ^* , $B \subseteq \Xi$, and a function $c(\cdot)$ from \mathcal{Y} to \mathbb{R}_+ , with $\mathbb{E}_q c(y_1) < \infty$, such that $\sup_{\xi \in B} \|H(y_1|\xi)\|_\infty \leq c(y_1)$;
8. the upper level sets of the average log-likelihood $\bar{\ell}_n(y^{(n)}|\cdot) = n^{-1} \sum_{i=1}^n \ell(y_i|\cdot)$, that is, all sets of the form $\{\xi \in \Xi : \bar{\ell}_n(y^{(n)}|\xi) > \lambda\}$, with $\lambda \in \mathbb{R}$, are connected, for all $y^{(n)} \in \mathcal{Y}^n$.

A singleton model $\mathcal{M} = \{f^n(\cdot|\xi^*)\}$ is regular with respect to $q^n(\cdot)$ if 2. and 4. hold, being understood that $K^* = K_q(\xi^*) < \infty$ and $\Xi = \{\xi^*\}$.

For a non-singleton model, the Kullback-Leibler divergence K^* from $q(\cdot)$ to \mathcal{M} is necessarily finite; otherwise ξ^* would not be uniquely defined. If $q(\cdot) \in \mathcal{M}$, then Assumption 3. is implied by identifiability, while Assumption 4. is trivial, because $f(\cdot|\xi^*) = q(\cdot)$. On the other hand, if $q(\cdot) \notin \mathcal{M}$, then $K^* > 0$, because $K_q(\xi) = 0$ implies $q(\cdot) = f(\cdot|\xi)$.

Assumption 7. in Definition A.1 implies $\mathbb{E}_q\|H(y_1|\xi)\|_\infty < \infty$ for all $\xi \in B$ and can be extended to the unit score vector by writing the Taylor expansion with integral remainder

$$s(y_1|\xi) = s(y_1|\xi^*) + \int_0^1 H(y_1|\xi^* + t(\xi - \xi^*))(\xi - \xi^*) dt,$$

which gives $\sup_{\xi \in B} \|s(y_1|\xi)\|_\infty \leq \|s(y_1|\xi^*)\|_2 + d c(y_1) \sup_{\xi \in B} \|\xi - \xi^*\|_2 = b(y_1)$; this in turn implies $\mathbb{E}_q\|s(y_1|\xi)\|_2 < \infty$ for all $\xi \in B$. Similarly, Assumption 7. in Definition A.1 can be extended to the unit log-likelihood, obtaining

$$\sup_{\xi \in B} |\ell(y_1, \xi)| \leq |\ell(y_1, \xi^*)| + d b(y_1) \sup_{\xi \in B} \|\xi - \xi^*\|_2 = a(y_1)$$

and $\mathbb{E}_q|\ell(y_1|\xi)| < \infty$ for all $\xi \in B$. We are now ready for a technical lemma.

LEMMA A.1. *If a non-singleton statistical model $\mathcal{M} = \{f^n(\cdot|\xi), \xi \in \Xi\}$ is regular with respect to the strictly positive sampling density $q^n(\cdot)$, the expected log-likelihood $L_q(\xi) = \mathbb{E}_q \ell(y_1, \xi)$, $\xi \in B$, is twice continuously differentiable on B with gradient $L'_q(\cdot) = \mathbb{E}_q s(y_1|\cdot)$ and Hessian matrix $L''_q(\cdot) = \mathbb{E}_q H(y_1|\cdot)$.*

PROOF. $L_q(\cdot)$ is differentiable on B with gradient $L'_q(\cdot) = \mathbb{E}_q s(y_1|\cdot)$ because Assumption 7. extended to the unit score vector allows the derivative to pass under the integral sign; see for instance the Lemma on page 124 of Ferguson (1996). In the same way, it follows from Assumption 7. that the expected score vector $L'_q(\cdot)$ is differentiable on B with derivative matrix $L''_q(\cdot) = \mathbb{E}_q H(y_1|\cdot)$. Then, through a direct application of Lebesgue's Dominated Convergence Theorem, it also follows from Assumption 7. that $L''_q(\cdot)$ is continuous on B . \square

Since $K_q(\xi) = K_q(\xi^*) + L_q(\xi^*) - L_q(\xi)$, for $\xi \in B$, with $K_q(\xi^*) < \infty$, the Kullback-Leibler divergence from $q(\cdot)$ to $f(\cdot, \xi)$ is also twice continuously differentiable on B , as a function of ξ , with gradient $K'_q(\cdot) = -L'_q(\cdot)$ and Hessian matrix $K''_q(\cdot) = -L''_q(\cdot)$. Then, since ξ^* is the unique minimum of $K_q(\cdot)$ on Ξ ,

ξ^* is also the unique maximum of $L_q(\cdot)$ on B . Finally, since ξ^* is an interior point of B , we find $L'_q(\xi^*) = K'_q(\xi^*) = 0$ and $L''_q(\xi^*)$ a negative definite matrix, equivalently, $K''_q(\xi^*)$ a positive definite matrix.

We now give a classical theorem on maximum likelihood asymptotics, but for a possibly mis-specified model; see Theorems 17 and 18 of Ferguson (1996).

THEOREM A.1. *Let $y^{(n)} = (y_1, \dots, y_n)$ be data arising under i.i.d. sampling from a distribution with strictly positive density $q^n(\cdot)$ and $\mathcal{M} = \{f^n(\cdot|\xi), \xi \in \Xi\}$ be a non-singleton model for such data, which we assume to be regular with respect to $q^n(\cdot)$. Then, there exists $\hat{\xi}_n$ such that, almost surely, for large enough n , $\hat{\xi}_n$ is a global maximum of the log-likelihood. Moreover, for any such maximum likelihood estimator $\hat{\xi}_n$, the following conditions are satisfied:*

- (i) almost surely, for large enough n , $\hat{\xi}_n$ is a root of the score equation, that is, $\bar{s}_n(y^{(n)}, \hat{\xi}_n) = 0$, where $\bar{s}_n(y^{(n)}, \cdot) = n^{-1} \sum_{i=1}^n s(y_i|\cdot)$ is the average score;
- (ii) almost surely $\hat{\xi}_n \rightarrow \xi^*$, as $n \rightarrow \infty$;
- (iii) for all small enough $\rho > 0$ there exists $\delta > 0$ such that, almost surely, for large enough n , it holds that

$$\sup_{\xi \in \Xi \cap \{\|\xi - \xi^*\|_2 \geq \rho\}} \bar{\ell}_n(y^{(n)}|\xi) < \bar{\ell}_n(y^{(n)}|\hat{\xi}_n) - \delta;$$

- (iv) $n^{1/2}(\hat{\xi}_n - \xi^*) \rightsquigarrow \mathcal{N}_d(0, V^*)$, as $n \rightarrow \infty$, where

$$V^* = L''_q(y_1|\xi^*)^{-1} \mathbb{E}_q\{s(y_1|\xi^*)s(y_1|\xi^*)^\top\} L''_q(y_1|\xi^*)^{-1}.$$

Finally, as a consequence of (iv), we can write $\hat{\xi}_n - \xi^* = \mathcal{O}_p(n^{-1/2})$.

PROOF. Let S be a compact sphere contained in B and denote by $\hat{\xi}_n$ a maximum likelihood estimator of ξ constrained to S , which always exists because the average log-likelihood $\bar{\ell}_n(y^{(n)}|\cdot)$ is continuous on S . Then, fix $\rho > 0$ small enough for $C = \{\xi \in S : \|\xi - \xi^*\|_2 \geq \rho\}$ to be non-empty. Since C is compact, a uniform version of the Strong Law of Large Numbers for continuous dominated summands (Ferguson, 1996, Theorem 16) gives $\sup_{\xi \in C} |\bar{\ell}_n(y^{(n)}|\xi) - L_q(\xi)| \rightarrow 0$, as $n \rightarrow \infty$, almost surely. Now $\sup_{\xi \in C} L_q(\xi) = L_q(\xi^*) - 3\delta$, for some $\delta > 0$, because $\xi^* \notin C$ and $L_q(\cdot)$ is continuous on C . Hence, almost surely, for large enough n , we have $\sup_{\xi \in C} \bar{\ell}_n(y^{(n)}|\xi) < L_q(\xi^*) - 2\delta$. However, due to the ordinary Strong Law of Large Numbers, we also have $\bar{\ell}_n(y^{(n)}|\xi^*) > L_q(\xi^*) - \delta$. Then, the connected set $\{\xi \in \Xi : \bar{\ell}_n(y^{(n)}|\xi) > L_q(\xi^*) - 2\delta\}$ contains ξ^* but has empty intersection with C . Since $C \neq \emptyset$, this upper level set has also empty intersection with $\Xi \setminus S$. It follows that $\hat{\xi}_n$ is a global maximum of the log-likelihood.

Now let $\hat{\xi}_n$ be any global maximum likelihood estimator. Since Ξ is open, $\hat{\xi}_n$ is necessarily an interior point of Ξ and (i) follows. Moreover, the above argument shows that $\|\hat{\xi}_n - \xi^*\|_2 < \rho$, for any small enough $\rho > 0$, which is enough to prove (ii). The above argument also gives

$$\sup_{\xi \in \Xi \cap \{\|\xi - \xi^*\|_2 \geq \rho\}} \bar{\ell}_n(y^{(n)}|\xi) \leq L_q(\xi^*) - 2\delta < \bar{\ell}_n(y^{(n)}|\hat{\xi}_n) - \delta,$$

which is (iii). Therefore, only (iv) remains to be shown.

Consider the Taylor expansion with integral remainder

$$\bar{s}_n(y^{(n)}|\xi^*) = \int_0^1 \bar{H}_n(y^{(n)}|\hat{\xi}_n + t(\xi^* - \hat{\xi}_n))(\xi^* - \hat{\xi}_n)dt,$$

where $\bar{H}_n(y^{(n)}|\cdot) = n^{-1} \sum_{i=1}^n H(y_i|\cdot)$ is the average Hessian of the log-likelihood and we have used the fact that $\hat{\xi}_n$ is a root of the score equation. The Central Limit Theorem tells us that $\bar{s}_n(y^{(n)}|\xi^*) \rightsquigarrow \mathcal{N}_d(0, \mathbb{E}_q\{s(y_1|\xi^*)s(y_1|\xi^*)^\top\})$, as $n \rightarrow \infty$. Hence, Slutsky's Theorem will give us (iv) if we show that

$$R_n(\hat{\xi}_n, \xi^*) = \int_0^1 \bar{H}_n(y^{(n)}|\hat{\xi}_n + t(\xi^* - \hat{\xi}_n))dt \rightarrow L_q''(\xi^*), \quad \text{as } n \rightarrow \infty,$$

almost surely, as we do below; notice that $L_q''(\xi^*)$ is negative definite and thus $R_n(\hat{\xi}_n, \xi^*)$ will be non-singular, for large enough n , almost surely.

Fix $\epsilon > 0$ and find $\rho > 0$ such that $\|L_q''(\xi) - L_q''(\xi^*)\|_\infty < \epsilon/2$ if $\|\xi - \xi^*\|_2 \leq \rho$; this is possible because $L_q''(\cdot)$ is continuous. Then, observe that $\|\hat{\xi}_n - \xi^*\|_2 \leq \rho$, for large enough n , almost surely, because of (ii). Therefore, we can write

$$\|R_n(\hat{\xi}_n, \xi^*) - L_q''(\xi^*)\|_\infty < \sup_{\xi: \|\xi - \xi^*\|_2 \leq \rho} \left\| \bar{H}_n(y^{(n)}|\xi) - L_q''(\xi) \right\|_\infty + \frac{\epsilon}{2},$$

for large enough n , almost surely, where the first term in the right hand side can be made smaller than $\epsilon/2$ by the same uniform Strong Law of Large Numbers invoked above. The thesis follows, because ϵ is arbitrary. \square

If $q(\cdot) \in \mathcal{M}$, the asymptotic covariance matrix V^* in Theorem A.1 is the inverse of Fisher's information matrix at ξ^* ; this can be shown through a well-known argument relying on passing the derivative under the integral sign (Ferguson, 1996, Chapter 18).

Next, in order to study the asymptotic behaviour of the marginal likelihood, we need to introduce the notion of a regular generalized moment prior.

DEFINITION A.2. *A generalized moment prior $p^M(\cdot) \propto g(\cdot)p(\cdot)$ on the open parameter space $\Xi \subseteq \mathfrak{R}^d$ is regular if the following assumptions hold:*

1. $p(\cdot)$ is a strictly positive continuous probability density on Ξ (local prior);
2. $g(\cdot)$ is an infinitely smooth function from Ξ to \mathfrak{R}_+ , whose k -th derivative we denote by $g^{(k)}(\cdot)$;
3. for all $\xi \in \Xi$ the least positive integer h such that $g^{(2h)}(\xi) \neq 0$ (order of the generalized moment prior at ξ) is finite.

It is intended that the normalizing constant $C_g = \int_{\Xi} g(\xi)p(\xi)d\xi$ be finite, as well as strictly positive, so that $p^M(\cdot)$ is a proper prior.

Notice that $g(\xi) = 0$ implies $g'(\xi) = 0$ and $g''(\xi)$ positive semidefinite, because $g(\cdot)$ is a function to \mathfrak{R}_+ . By iterating this argument, we find that $g^{(2h-1)}(\xi) = 0$ and $g^{(2h)}(\xi)$ is a positive semidefinite, non-null, multilinear form on \mathfrak{R}^{2h} .

We are now ready to give our main result on the marginal likelihood of a regular model with regular generalized moment prior.

THEOREM A.2. Let $y^{(n)} = (y_1, \dots, y_n)$ be data arising under i.i.d. sampling from an unknown distribution with strictly positive density $q^n(\cdot)$ and $\mathcal{M} = \{f^n(\cdot|\xi), \xi \in \Xi\}$ be a non-singleton statistical model for such data, which we assume to be regular with respect to $q^n(\cdot)$. Denote by $m^M(y^{(n)})$ the marginal likelihood of \mathcal{M} under a regular generalized moment prior $p^M(\cdot)$. Then:

(i) if $q(\cdot) \notin \mathcal{M}$,

$$\log \frac{m^M(y^{(n)})}{q^n(y^{(n)})} = -nK^* + \mathcal{O}_p(n^{1/2});$$

(ii) if $q(\cdot) \in \mathcal{M}$,

$$\log \frac{m^M(y^{(n)})}{q^n(y^{(n)})} = -\frac{d}{2} \log n - h^* \log n + \mathcal{O}_p(1),$$

where h^* is the order of $p^M(\cdot)$ at ξ^* .

If \mathcal{M} is a singleton model, which needs no prior, then (i) holds unchanged and (ii) holds trivially with $d = 0$ and $h^* = 0$.

PROOF. Following Dawid (2011) we factorize the ratio of the marginal likelihood to the unknown sampling distribution as

$$(24) \quad \frac{m^M(y^{(n)})}{q^n(y^{(n)})} = \frac{m^M(y^{(n)})}{f^n(y^{(n)}|\hat{\xi}_n)} \times \frac{f^n(y^{(n)}|\hat{\xi}_n)}{f^n(y^{(n)}|\xi^*)} \times \frac{f^n(y^{(n)}|\xi^*)}{q^n(y^{(n)})}.$$

We deal with the three factors, which we name F_1 , F_2 and F_3 , in reverse order. Notice that F_1 and F_2 are (to be considered) identically one for a singleton model.

The third factor in (24) is trivially one if $q(\cdot) \in \mathcal{M}$, because in this case $f(\cdot|\xi^*) = q(\cdot)$. On the other hand, if $q(\cdot) \notin \mathcal{M}$, its logarithm can be written as

$$\log F_3 = \sum_{i=1}^n \log \frac{f(y_i|\xi^*)}{q(y_i)},$$

that is, as a sum of i.i.d. random numbers with expectation $-K^*$. It follows from the Central Limit Theorem that

$$\frac{1}{\sqrt{n}}(\log F_3 + nK^*) \rightsquigarrow \mathcal{N}_1 \left(0, \mathbb{V}_q \log \frac{q(y_1)}{f(y_1|\xi^*)} \right), \quad \text{as } n \rightarrow \infty,$$

and thus we find $\log F_3 = -nK^* + \mathcal{O}_p(n^{1/2})$.

The logarithm of the second factor in (24) can be written as

$$\log F_2 = -n \int_0^1 (1-u)(\xi^* - \hat{\xi}_n)^\top \bar{H}_n(y^{(n)}|\hat{\xi}_n + u(\xi^* - \hat{\xi}_n))(\xi^* - \hat{\xi}_n) du,$$

using a Taylor expansion with integral reminder of the average log-likelihood about $\hat{\xi}_n$; remember that $\hat{\xi}_n$ is a root of the score equation. Like in the proof of Theorem A.1, it can be shown that

$$\int_0^1 (1-u) \bar{H}_n(y^{(n)}|\hat{\xi}_n + u(\xi^* - \hat{\xi}_n)) du = \frac{1}{2} L_q''(\xi^*) + o_p(1).$$

Then, since we know from Theorem A.1 that $n^{1/2}(\hat{\xi}_n - \xi^*) = \mathcal{O}_p(1)$, we find $\log F_2 = \mathcal{O}_p(1)$; this holds regardless of $q(\cdot) \in \mathcal{M}$ or $q(\cdot) \notin \mathcal{M}$.

The first factor in (24) can be dealt with by means of a Laplace approximation of the marginal likelihood. Specifically, for all sufficiently small $\rho > 0$, the latter can be written as $m^M(y^{(n)}) = I_\rho(y^{(n)}) + I_\rho^*(y^{(n)})$, where

$$\begin{aligned} I_\rho(y^{(n)}) &= \int_{\{\xi \in \Xi: \|\xi - \xi^*\|_2 > \rho\}} f^n(y^{(n)}|\xi) p^M(\xi) d\xi, \\ I_\rho^*(y^{(n)}) &= \int_{\{\xi \in \mathbb{R}^d: \|\xi - \xi^*\|_2 \leq \rho\}} f^n(y^{(n)}|\xi) p^M(\xi) d\xi. \end{aligned}$$

By (iii) of Theorem A.1 we can find $\delta > 0$ such that, almost surely, for large enough n , $I_\rho(y^{(n)}) \leq f^n(y^{(n)}|\hat{\xi}_n) e^{-\delta n}$; it follows that, by increasing n , we can make $I_\rho(y^{(n)})/\{n^{-h^* - d/2} f^n(y^{(n)}|\hat{\xi}_n)\}$ as small as we like. On the other hand, a Taylor expansion of the average log-likelihood about $\hat{\xi}_n$ gives us

$$\frac{I_\rho^*(y^{(n)})}{f^n(y^{(n)}|\hat{\xi}_n)} = \int_{\{\xi \in \mathbb{R}^d: \|\xi - \xi^*\|_2 \leq \rho\}} \exp \left\{ -n(\xi - \hat{\xi}_n)^\top R_n(\xi, \hat{\xi}_n)(\xi - \hat{\xi}_n) \right\} p^M(\xi) d\xi,$$

where $R_n(\xi, \hat{\xi}_n) = -\int_0^1 (1-u) \bar{H}_n(y^{(n)}|\hat{\xi}_n + u(\xi - \hat{\xi}_n)) du$ is the integral reminder. Now fix $\epsilon > 0$. Like in the proof of Theorem A.1, a suitable choice of ρ makes $\|R_n(\xi, \hat{\xi}_n) - \frac{1}{2} K_q''(\xi^*)\|_\infty$ smaller than ϵ , for large enough n , almost surely. In this way, we obtain $J_n(-\epsilon) \leq I_\rho^*(y^{(n)})/f^n(y^{(n)}|\hat{\xi}_n) \leq J_n(\epsilon)$, where

$$J_n(\epsilon) = \int_{\{\xi \in \mathbb{R}^d: \|\xi - \xi^*\|_2 \leq \rho\}} \exp \left\{ -\frac{n}{2} (\xi - \hat{\xi}_n)^\top (K_q''(\xi^*) - 2\epsilon d^2 I_d) (\xi - \hat{\xi}_n) \right\} p^M(\xi) d\xi,$$

with I_d denoting the $d \times d$ identity matrix. In the following we deal with $J_n(-\epsilon)$ implicitly, by considering $J_n(\epsilon)$ without assuming $\epsilon > 0$.

Since $p(\xi) = p(\xi^*) + o(1)$ and $g(\xi) = \frac{1}{(2h^*)!} g^{(2h^*)}(\xi^*) [(\xi - \xi^*)^{2h^*}] + o(\|\xi - \xi^*\|_2^{2h^*})$, as $\xi \rightarrow \xi^*$, we have

$$p^M(\xi) = \frac{p(\xi^*)}{C_g(2h^*)!} g^{(2h^*)}(\xi^*) [(\xi - \xi^*)^{2h^*}] + o(\|\xi - \xi^*\|_2^{2h^*}), \quad \text{as } \xi \rightarrow \xi^*,$$

and a suitable choice of ρ makes $|J_n(\epsilon) - \{C_g(2h^*)!\}^{-1} p(\xi^*) J_n^*(\epsilon)|/\tilde{J}_n(\epsilon)$ as small as we like, where

$$J_n^*(\epsilon) = \int_{\{\xi \in \mathbb{R}^d: \|\xi - \xi^*\|_2 \leq \rho\}} \exp \left\{ -\frac{n}{2} (\xi - \hat{\xi}_n)^\top \Lambda_\epsilon (\xi - \hat{\xi}_n) \right\} g^{(2h^*)}(\xi^*) [(\xi - \xi^*)^{2h^*}] d\xi$$

and

$$\tilde{J}_n(\epsilon) = \int_{\{\xi \in \mathbb{R}^d: \|\xi - \xi^*\|_2 \leq \rho\}} \exp \left\{ -\frac{n}{2} (\xi - \hat{\xi}_n)^\top \Lambda_\epsilon (\xi - \hat{\xi}_n) \right\} \|\xi - \xi^*\|_2^{2h^*} d\xi,$$

with $\Lambda_\epsilon = K_q''(\xi^*) - 2\epsilon d^2 I_d$. Both $J_n^*(\epsilon)$ and $\tilde{J}_n(\epsilon)$ are of the form

$$J_n^A(\epsilon) = \int_{\{\xi \in \mathbb{R}^d: \|\xi - \xi^*\|_2 \leq \rho\}} \exp \left\{ -\frac{n}{2} (\xi - \hat{\xi}_n)^\top \Lambda_\epsilon (\xi - \hat{\xi}_n) \right\} A [(\xi - \xi^*)^{2h^*}] d\xi,$$

where A is a positive semidefinite, non-null, multilinear form on \mathfrak{R}^{2h} . We consider below the extension of $J_n^A(\epsilon)$ to \mathfrak{R}^d , which we denote by $\bar{J}_n^A(\epsilon)$.

By writing $A[(\xi - \xi^*)^{2h^*}] = \sum_{i=0}^{2h^*} \binom{2h^*}{i} A[(\hat{\xi}_n - \xi^*)^i (\xi - \hat{\xi}_n)^{2h^*-i}]$, and operating the change of variable $\zeta = n^{1/2}(\xi - \hat{\xi}_n)$, we obtain

$$\begin{aligned} \bar{J}_n^A(\epsilon) &= \int_{\mathfrak{R}^d} \exp\left\{-\frac{n}{2}(\xi - \hat{\xi}_n)^\top \Lambda_\epsilon (\xi - \hat{\xi}_n)\right\} A[(\xi - \xi^*)^{2h^*}] d\xi \\ &= \sum_{i=0}^{2h^*} \binom{2h^*}{i} \int_{\mathfrak{R}^d} \exp\left\{-\frac{1}{2}\zeta^\top \Lambda_\epsilon \zeta\right\} A[\{n^{1/2}(\hat{\xi}_n - \xi^*)\}^i \zeta^{2h^*-i}] n^{-h^* - \frac{d}{2}} d\zeta \\ &= \frac{(2\pi)^{\frac{d}{2}}}{|\Lambda_\epsilon|^{\frac{1}{2}}} n^{-h^* - d/2} \mathbb{E}A[Z_\epsilon^{2h^*}] \{1 + \mathcal{O}_p^+(1)\}, \end{aligned}$$

where Z_ϵ is a normal random vector with zero mean and precision matrix Λ_ϵ , and $\mathcal{O}_p^+(1)$ is a positive $\mathcal{O}_p(1)$ term; positivity follows from even values of i giving positive terms, and odd values of i giving zero terms, in the above displayed sum. We are now ready to conclude our proof.

Since $|\bar{J}_n^A(\epsilon) - J_n^A(\epsilon)|$ is less than

$$\int_{\{\xi \in \mathfrak{R}^d: \|\xi - \hat{\xi}_n\|_2 > \rho/2\}} \exp\left\{-\frac{n}{2}(\xi - \hat{\xi}_n)^\top \Lambda_\epsilon (\xi - \hat{\xi}_n)\right\} A[(\xi - \xi^*)^{2h^*}] d\xi,$$

if n is large enough to have $\|\hat{\xi}_n - \xi^*\|_2 < \rho/2$, the same computations carried out above show that $|\bar{J}_n^A(\epsilon) - J_n^A(\epsilon)|/n^{-h^* - d/2}$ is arbitrarily small, for large enough n . Hence, we have $J_n^A(\epsilon) = (2\pi)^{d/2} |\Lambda_\epsilon|^{-1/2} n^{-h^* - d/2} \mathbb{E}A[Z_\epsilon^{2h^*}] \{1 + \mathcal{O}_p^+(1)\}$ and then $J_n(\epsilon) = \{C_g(2h^*)!\}^{-1} p(\xi^*) (2\pi)^{d/2} |\Lambda_\epsilon|^{-1/2} n^{-h^* - d/2} \mathbb{E}g^{(2h^*)}(\xi^*) [Z_\epsilon^{2h^*}] \{1 + \mathcal{O}_p^+(1)\}$. Finally, since ϵ is arbitrary, it follows that

$$\frac{I_\rho^*(y^{(n)})}{f^n(y^{(n)}|\hat{\xi}_n)} = \frac{p(\xi^*)}{C_g(2h^*)!} \frac{(2\pi)^{\frac{d}{2}}}{|K_q''(\xi^*)|^{\frac{1}{2}}} \mathbb{E}\{g^{(2h^*)}(\xi^*) [Z^{2h^*}]\} n^{-h^* - \frac{d}{2}} \{1 + \mathcal{O}_p^+(1)\},$$

where Z is a normal random vector with zero mean and precision matrix $K_q''(\xi^*)$; this eventually leads to $\log F_1 = -h^* \log n - \frac{d}{2} \log n + \mathcal{O}_p(1)$ as desired. \square

The above theorem also covers local priors, by letting $g(\cdot) \equiv 1$, so that h^* is identically zero; in this case it essentially returns the result of Dawid (2011).

We are now in a position to describe the asymptotic behaviour of BFs using a generalized moment prior under the alternative and a local prior under the null (or comparing to a point null).

COROLLARY A.1. *Let $\mathcal{M}_0 \subset \mathcal{M}_1$ be two nested models for the same data $y^{(n)} = (y_1, \dots, y_n)$ and assume that both these models are regular with respect to all distributions in \mathcal{M}_1 , with dimensions $d_0 < d_1$. Denote by $BF_{10}^M(y^{(n)})$ the Bayes factor in favour of \mathcal{M}_1 against \mathcal{M}_0 using a regular generalized moment prior under \mathcal{M}_1 with order h on the subspace of \mathcal{M}_1 corresponding to \mathcal{M}_0 . If \mathcal{M}_0 is a non-singleton model, let it be equipped with a local prior. Finally, denote by $q^n(\cdot)$ the actual sampling distribution and recall that $BF_{01}^M(y^{(n)}) = 1/BF_{10}^M(y^{(n)})$. Then:*

(i) if $q(\cdot) \in \mathcal{M}_1 \setminus \mathcal{M}_0$,

$$BF_{01}(y^{(n)}) = \exp\{-nK^* + \mathcal{O}_p(n^{1/2})\};$$

(ii) if $q(\cdot) \in \mathcal{M}_0$,

$$BF_{10}(y^{(n)}) = \exp\left\{-\frac{(d_1 - d_0)}{2} \log n - h \log n + \mathcal{O}_p(1)\right\}.$$

PROOF. This follows directly from Theorem A.2. □