UNIVERSITY OF Southampton
Southampton Statistical Sciences Research Institute

Working Paper M10/13
Methodology

Enhanced Objective Bayesian Testing For The Equality Of Two Proportions

Guido Consonni, Jonathan J. Forster, Luca La Rocca

Abstract

We develop a new class of prior distributions for Bayesian comparison of nested models, which we call intrinsic moment priors. We aim at testing the equality of two proportions, based on independent samples, and thus focus on discrete data models. We illustrate our methodology in a running Bernoulli example, where we test a sharp null hypothesis, then we implement our procedure to test the equality of two proportions. A detailed analysis of the properties of our method is presented together with an application to a collection of real-world 2*2 tables involving a sensitivity analysis and a crossvalidation study.

Enhanced Objective Bayesian Testing for the Equality of two Proportions

Guido Consonni, Jonathan J. Forster and Luca La Rocca [1]

October 15, 2010

[1]Guido Consonni is Professor of Statistics, Dipartimento di Economia Politica e Metodi Quantitativi, University of Pavia, Pavia, Italy (email: guido.consonni@unipv.it); Jon Forster is Professor of Statistics, School of Mathematics, University of Southampton, Southampton, UK (email: j.j.forster@soton.ac.uk); and Luca La Rocca is Assistant Professor of Statistics, Dipartimento di Comunicazione e Economia, University of Modena and Reggio Emilia, Reggio Emilia, Italy (email: luca.larocca@unimore.it). This work was partially supported by MIUR, Rome, PRIN 2007XECZ7L_001, and the University of Pavia; it was mostly developed 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.

Abstract

We develop a new class of prior distributions for Bayesian comparison of nested models, which we call *intrinsic moment* priors, by combining the well-established notion of *intrinsic* prior with the recently introduced idea of *non-local* priors, and in particular of *moment* priors. Specifically, we aim at testing the equality of two proportions, based on independent samples, and thus focus on discrete data models. Given two nested models, each equipped with a default prior, we first construct a moment prior under the larger model. In this way, the asymptotic learning behavior of the Bayes factor is strengthened, relative to currently used *local* priors, when the smaller model holds; remarkably, this effect is already apparent for moderate sample sizes. On the other hand, the asymptotic learning behavior of the Bayes factor when the larger model holds is unchanged. However, without appropriate tuning, a moment prior does not provide enough evidence for the larger model when the sample size is small and the data only moderately support the smaller one. For this reason, we apply to the moment prior an intrinsic prior procedure, which amounts to pulling the moment prior towards the subspace specified by the smaller model; we provide general guidelines for determining the training sample size necessary to implement this step. Thus, by joining the virtues of moment and intrinsic priors, we obtain an *enhanced* objective Bayesian testing procedure: i) our evidence for small samples is broadly comparable to that given by current objective methods; ii) we achieve a superior learning performance as the sample size increases (when the smaller model holds). We first illustrate our methodology in a running Bernoulli example, where we test a sharp null hypothesis, then we implement our procedure to test the equality of two proportions. A detailed analysis of the properties of our method, including a comparison with standard intrinsic priors, is presented together with an application to a collection of real-world 2×2 tables involving a sensitivity analysis and a cross-validation study.

Keywords: Bayes factor; intrinsic prior; model choice; moment prior; non-local prior; training sample size.

1 Introduction

The analysis of two independent binomial populations has a long history, dating back to the beginning of the twentieth century; for an account see Howard (1998). If θ_1 and θ_2 denote the two population proportions, a typical hypothesis of interest is that of equality, $\theta_1 = \theta_2$, or a one sided hypothesis, such as $\theta_1 < \theta_2$. For the latter case Howard (1998) provides a Bayesian reinterpretation of frequentist tests as well as a proposal for a fully Bayesian analysis based on a prior distribution embodying dependence between θ_1 and θ_2.

In this paper we are concerned with testing the equality of the two proportions. The problem is equivalent to testing independence in a 2×2 contingency table whose row margins are fixed by design. Casella and Moreno (2009) recently provided a detailed objective Bayesian analysis of the latter problem, also discussing alternative sampling procedures, based on the notion of *intrinsic* prior.

Intrinsic priors are now recognized as a useful tool for Bayesian hypothesis testing and more generally for model comparison, especially in an objective Bayesian setting. Numerous applications ranging from variable selection (Casella and Moreno, 2006; Casella et al., 2009) to change point problems (Moreno et al., 2005; Girón et al., 2007) to contingency tables (Casella and Moreno, 2005; Consonni and La Rocca, 2008; Casella and Moreno, 2009) testify their potential.

Given two parametric models, $\mathcal{M}_0$ (the *null* model) nested in $\mathcal{M}_1$ (the *alternative* model), each equipped with its own default prior distribution, $p_0^D(\cdot)$ and $p_1^D(\cdot)$, the intrinsic (prior) approach suitably modifies $p_1^D(\cdot)$ by "peaking" it around the subspace specified by $\mathcal{M}_0$. In this way, $\mathcal{M}_1$ becomes more competitive against $\mathcal{M}_0$ precisely when the comparison is most delicate, namely for data generating mechanisms close to the null, and this displacement of prior mass effectively averts the Jeffreys-Lindley-Bartlett paradox; see Kass and Raftery (1995, sect. 5.1) and Robert (2001, sect. 5.2.5).

The idea of "centering" the prior around the null, under the larger model, can be traced back at least to Jeffreys and can also be performed outside the intrinsic prior setup. A notable example in this sense is the hierarchical Bayesian framework, as developed by Albert and Gupta (1982) and Albert (1990).

Virtually all priors under $\mathcal{M}_1$ currently used for Bayesian hypothesis testing or model comparison belong to the class of *local* priors, which do *not* vanish over the subspace specified by the null. For instance, in the problem of testing the equality of two proportions, the default prior will be a product of uniform priors (one for θ_1 and one for θ_2), or possibly a product of two Jeffreys priors. Clearly, these priors are bounded away from zero on the line $\theta_1 = \theta_2$. The intrinsic prior for this problem shares a similar feature; see Casella and Moreno (2009, sect. 3.2).

A serious deficiency of local priors relates to their asymptotic learning rate. Specifically, the Bayes factor in favor of $\mathcal{M}_1$, when $\mathcal{M}_1$ holds, diverges in probability exponentially fast, as the sample size grows, whereas it converges to zero in probability at polynomial rate only, when $\mathcal{M}_0$ holds. Although this fact is well known, it is less known that this imbalance is already quite dramatic for moderate sample sizes. However, this feature can be successfully corrected, as suggested in recent work of Johnson and Rossell (2010), where these authors advocate the use of *non-local* priors, and in particular of *moment* priors.

We find the idea of non-local priors appealing. At the same time, we concur that the rationale underlying the intrinsic approach for tuning a default prior is useful. We therefore combine non-local and intrinsic priors into a *unified new class of priors for testing nested models.* These priors exhibit finite sample properties of the Bayes factor comparable to those of the intrinsic approach, but they outperform current local prior approaches (including the intrinsic one) in terms of asymptotic learning behavior (when the null model holds). In this sense, we obtain an *enhanced* Bayesian testing procedure.

The rest of the paper is organized as follows. Section 2 provides background material on intrinsic and moment priors, with special reference to the testing problem under consideration. Sections 3 and 4 represent the core of the paper: the former presents a new class of non-local priors, which we name *intrinsic moment* priors, while the latter implements the proposed methodology to obtain an enhanced objective Bayesian test for the equality of two proportions. Section 5 applies our new test 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). Section 6 offers some concluding remarks and investigates a few issues worth of further consideration.

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 moment priors.

Consider two sampling models for the same *discrete* observables:

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

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 $\xi_1 \in \tilde{\Xi}_0 \subset \Xi_1$. Let $p_0(\xi_0)$ and $p_1(\xi_1)$ be the priors under the two models, which we assume proper, and denote the data by $y = (y_1, \ldots, y_n)$; occasionally we will write $y^{(n)}$ to stress the dependence on n. The Bayes factor in *favor* of $\mathcal{M}_1$ (equivalently *against* $\mathcal{M}_0$) is $BF_{10}(y) = \frac{m_1(y)}{m_0(y)}$, where $m_j(y) = \int f_j(y|\xi_j)p_j(\xi_j)\mathrm{d}\xi_j$, $j = 0, 1$. We assume equal prior probabilities for $\mathcal{M}_0$ and $\mathcal{M}_1$, so that the posterior probability of $\mathcal{M}_1$ can be immediately recovered from $BF_{10}(y)$ as $\mathbb{P}(\mathcal{M}_1|y) = (1+BF_{01}(y))^{-1}$, where $BF_{01}(y) = 1/BF_{10}(y)$.

2.1 Intrinsic priors

Intrinsic priors were introduced in objective hypothesis testing to convert improper priors into proper ones (Berger and Pericchi, 1996; Moreno, 1997; Moreno et al., 1998). In this way, Bayes factors, which cannot be meaningfully evaluated using improper priors, admit a sensible interpretation. However, this view of the intrinsic approach is unduly restrictive and actually hinders its inherent nature, as it is apparent for discrete data models: in this case the default priors are usually proper, but the intrinsic approach may still be considered useful.

The actual implication of intrinsic priors is to "peak" the prior under the alternative around the region specified by the null, a suggestion dating back to Jeffreys; see also Morris (1987). This is related to the Jeffreys-Lindley-Bartlett paradox, because the idea is to counterbalance the excessive diffuseness of many standard default priors under the alternative. Casella and Moreno (2005), Consonni and La Rocca (2008) and Casella and Moreno (2009) reiterate this concept for discrete data models.

Let $p_0^D(\xi_0)$ and $p_1^D(\xi_1)$ be two *default* priors under $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively; for simplicity we assume them to be *proper*, as this will typically be the case with discrete data models. While we may in general retain $p_0^D(\xi_0)$, it is often the case that $p_1^D(\xi_1)$ be inappropriate because it is relatively too diffuse, thus unduly penalizing $\mathcal{M}_1$ when the data only mildly support $\mathcal{M}_0$. Let $x = (x_1, \ldots, x_t)$ be a vector of observables, whose dimensionality t we call the *training sample size.* The intrinsic prior on ξ_1 with training sample size t is given by

$$p_1^I(\xi_1|t) = \sum_x p_1^D(\xi_1|x) m_0^D(x), \tag{2}$$

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

We remark that (2) is not the original definition of intrinsic prior, but rather its formulation as an expected posterior prior (Perez and Berger, 2002). We find formula (2) especially appealing, because it makes clear that an intrinsic prior is a mixture of fictitious posteriors. Notice that, as the training sample size t increases, the intrinsic prior tends to "peak" on the subspace $\tilde{\Xi}_0$. The choice of t is left to the user, and it should be noticed that the standard notion of *minimal* training sample size is vacuous in the context of discrete observables, because the default priors are already proper by assumption.

Example 2.1 (Bernoulli) *Denoting ξ_1 by θ and ξ_0 by θ_0, consider the testing problem $\mathcal{M}_0 : f_0(y|\theta_0) = \mathrm{Bin}(y|n, \theta_0)$ versus $\mathcal{M}_1 : f_1(y|\theta) = \mathrm{Bin}(y|n, \theta)$, where θ_0 is a* fixed *value, while θ varies in $(0, 1)$. Let the default prior be $p_1^D(\theta|b) = \mathrm{Beta}(\theta|b, b)$ for some $b > 0$. We take a* symmetric *prior because standard default objective priors satisfy this property. The intrinsic prior in this case is given by*

$$p_1^I(\theta|b, t) = \sum_{x=0}^{t} \mathrm{Beta}(\theta|b + x, b + t - x) \mathrm{Bin}(x|n, \theta_0). \tag{3}$$

The solid curves in Figure 1(a), i.e., those specified by $h = 0$, illustrate the behavior of the intrinsic priors with training sample size $t = 0$ (default prior), $t = 1$ and $t = 8$, when $\theta_0 = 0.25$ and $b = 1$ (uniform default prior). The dashed curves ($h = 1$) should be discarded 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), for $n = 12$ and different values of the observed count y, the posterior probability of $\mathcal{M}_1$ as a function of the observed count y (evidence curve): switching from $t = 0$ (default prior) to $t = 1$ (intrinsic prior with unit training sample), the evidence in favor of the alternative model increases, especially in the region around $y = 3$, while correctly remaining well below the* 0.5*-line.*

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 Bayes factor for commonly used priors.

Result 2.1 *In the testing problem (1) assume that $p_1(\xi_1)$ is continuous and strictly positive for all $\xi_1 \in \Xi_1$, and let the data $y^{(n)} = (y_1, \ldots, y_n)$ arise under i.i.d. sampling. If $\mathcal{M}_0$ holds, then $BF_{10}(y^{(n)}) = O_p(n^{-(d_1-d_0)/2})$, as $n \to \infty$, where d_j is the dimension of Ξ_j, $j = 1, 2$, $(d_1 > d_0)$; if $\mathcal{M}_1$ holds, then $BF_{01}(y^{(n)}) = e^{-Kn+O_p(n^{1/2})}$, as $n \to \infty$, for some $K > 0$.*

We refer to Dawid (1999) 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 $\xi_1 \in \tilde{\Xi}_0$, so that 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 $\tilde{\Xi}_0$. This is indeed the approach taken by Johnson and Rossell (2010) when defining their non-local priors. Their motivation, however, is also conceptual, as it relates to an idea of separation between models.

Let $g(\xi_1)$ be a continuous function vanishing on $\tilde{\Xi}_0$. From a given *baseline* (local) prior $p_1(\xi_1)$, define a new prior as

$$p_1^M(\xi_1) \propto g(\xi_1) p_1(\xi_1), \tag{4}$$

which we name a *generalized moment* prior. For instance, if $\Xi_1 \subseteq \mathbb{R}$ and $\tilde{\Xi}_0 = \Xi_0 = \{\xi_0\}$, with ξ_0 a fixed value, we may take $g(\xi_1) = (\xi_1 - \xi_0)^{2h}$, where h is a positive

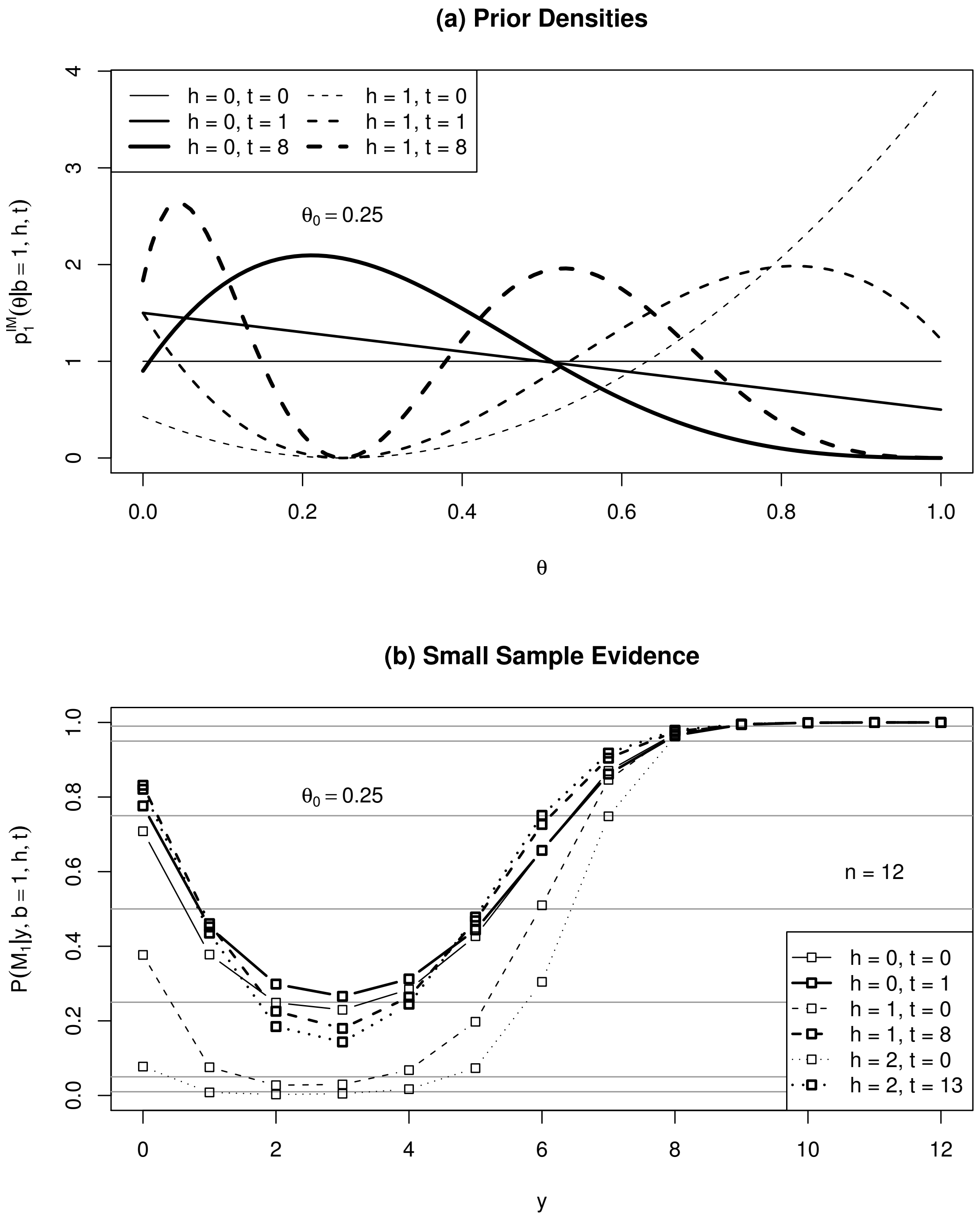

Figure 1: Prior densities and small sample evidence for the Bernoulli example.

integer ($h = 0$ returns the baseline prior); this defines the moment prior introduced by Johnson and Rossell (2010) for testing a sharp hypothesis on a scalar parameter. It can be seen that in this case $BF_{10}(y^{(n)}) = O_p(n^{-1/2-h})$ when $\mathcal{M}_0$ holds, while we still have $BF_{01}(y^{(n)}) = e^{-Kn+O_p(n^{1/2})}$ when $\mathcal{M}_1$ holds.

Example 2.2 (Bernoulli ctd.) *For a given conjugate baseline prior $Beta(\theta|a_1, a_2)$ define the corresponding* conjugate moment *prior of order h as*

$$p_1^{CM}(\theta|a_1, a_2, h) = \frac{(\theta - \theta_0)^{2h}}{K(a_1, a_2, h, \theta_0)} \times \mathrm{Beta}(\theta|a_1, a_2), \tag{5}$$

where

$$K(a_1, a_2, h, \theta_0) = \frac{\theta_0^{2h}}{B(a_1, a_2)} \sum_{j=0}^{2h} \binom{2h}{j} (-1)^j \theta_0^{-j} B(a_1 + j, a_2), \tag{6}$$

and $B(a_1, a_2)$ is the Beta function with parameters (a_1, a_2). In particular, if $h = 1$ and $a_1 = a_2 = b$, with b a default choice (such as $b = 1$ or $b = 1/2$), we obtain the default moment *prior $p_1^{DM}(\theta|b, h)$. The thin dashed curve in Figure 1(a), i.e., the one specified by $h = 1$ and $t = 0$, represents the default moment prior with $b = 1$ when $\theta_0 = 0.25$. The other two dashed curves ($h = 1$) with $t = 1$ (intermediate) and $t = 8$ (thick) should be ignored for the time being. The behavior of $p_1^{DM}(\theta|b = 1, h = 1)$ is the following: 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 figure shows that things get even worse for $h = 2$.*

The next section takes up the above issue and provides an effective solution.

3 Intrinsic Moment Priors

From Example 2.2 it is clear that 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, thus obtaining a new class of priors for testing nested hypotheses, which we name *intrinsic moment* priors.

Our strategy for enhanced objective Bayesian testing of two nested models thus envisages the following steps: i) start with a default prior under each of the two models; ii) construct the default moment prior of order h under the larger model; iii) for a given training sample size t, generate the corresponding intrinsic prior, which produces the intrinsic moment prior: this is the prior we recommend to compute the Bayes factor. Step ii) improves the learning behavior under the null, while step iii) makes sure that the testing procedure exhibits a good finite sample behavior in terms of the evidence curve .

Example 3.1 (Bernoulli ctd.) *Recall that the intrinsic prior is an average of fictitious posterior distributions. Since in our case we start from the default moment prior (5) with $a_1 = a_2 = b$, the intrinsic moment prior for θ with training sample size t will be given by*

$$p_1^{IM}(\theta|b,h,t) = \sum_{x=0}^{t} \frac{(\theta-\theta_0)^{2h}}{K(b+x, b+t-x, h, \theta_0)} \mathrm{Beta}(\theta|b+x, b+t-x)\mathrm{Bin}(x|t,\theta_0), \quad (7)$$

where $K(a_1, a_2, h, \theta_0)$ is defined in (6), and we exploited conjugacy of $p_1^{CM}(\theta|a_1, a_2, h)$. Notice that (7) describes a family of prior distributions, including *standard intrinsic priors ($h = 0$) as well as the default prior ($h = 0, t = 0$) as special cases. In order to compute the Bayes factor $BF_{10}^{IM}(y|b,h,t) = \frac{m_1^{IM}(y|b,h,t)}{f_0(y|\theta_0)}$, where $m_1^{IM}(y|b,h,t) = \int f_1(y|\theta)p_1^{IM}(\theta|b,h,t)\mathrm{d}\theta$ and $f_0(y|\theta_0) = \mathrm{Bin}(y|n,\theta_0)$, we exploit the fact that the Bayes factor using an intrinsic prior is a weighted average of conditional Bayes factors based on the starting prior; see for example Consonni and La Rocca (2008, Proposition 3.4). Thus, we find*

$$BF_{10}^{IM}(y|b,h,t) = \sum_{x=0}^{t} BF_{10}^{CM}(y|b+x, b+t-x, h)\mathrm{Bin}(x|t,\theta_0), \quad (8)$$

where $BF_{10}^{CM}(y|a_1,a_2,h) = \frac{m_1^{CM}(y|a_1,a_2,h)}{f_0(y|\theta_0)}$, and we then compute $m_1^{CM}(y|a_1,a_2,h) = \int f_1(y|\theta)p_1^{CM}(\theta|a_1,a_2,h)\mathrm{d}\theta$ by means of the useful relationship

$$m_1^{CM}(y|a_1,a_2,h) = \frac{K(a_1+y, a_2+n-y, h, \theta_0)}{K(a_1,a_2,h,\theta_0)} m_1^{C}(y|a_1,a_2), \quad (9)$$

where $m_1^C(y|a_1,a_2) = \int f_1(y|\theta)p_1^C(\theta|a_1,a_2)\mathrm{d}\theta = \binom{n}{y}\frac{B(a_1+y,a_2+n-y)}{B(a_1,a_2)}$ is the usual Beta-Binomial marginal density using the conjugate prior $p_1^C(\theta|a_1,a_2) = \mathrm{Beta}(\theta|a_1,a_2)$.

Notice that equation (9) reveals a structural relationship between the marginal data distribution based on a conjugate moment prior and that based on its conjugate baseline prior; its scope is in fact general. A consequence of (9) is that the Bayes factor based on a conjugate moment prior can be readily computed from the usual Bayes factor based on a conjugate prior:

$$BF_{10}^{CM}(y|a_1, a_2, h) = \frac{K(a_1 + y, a_2 + n - y, h, \theta_0)}{K(a_1, a_2, h, \theta_0)} BF_{10}^{C}(y|a_1, a_2), \tag{10}$$

where $BF_{10}^{C}(y|a_1, a_2) = \frac{m_1^C(y|a_1,a_2)}{f_0(y|\theta_0)}$.

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 count* y*; this intrinsic moment prior results in a smaller amount of evidence for values of* y *close to the null, which is to be expected for continuity, but induces a steeper evidence gradient as* y *moves away from the null, which makes it appealing.*

The learning behavior of the intrinsic moment prior is illustrated in Figure 2(a), which reports the average posterior probability of the null model (computed on 1000 simulated data sets of increasing size) letting first $\theta = 0.25$ *and then* $\theta = 0.4$ *(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 method actually represents a marked improvement over current methods. Notice that there is an associated cost: the moment prior trades off speed in learning the alternative model for speed in learning the null model; the intrinsic procedure is remarkably effective in controlling this trade off.*

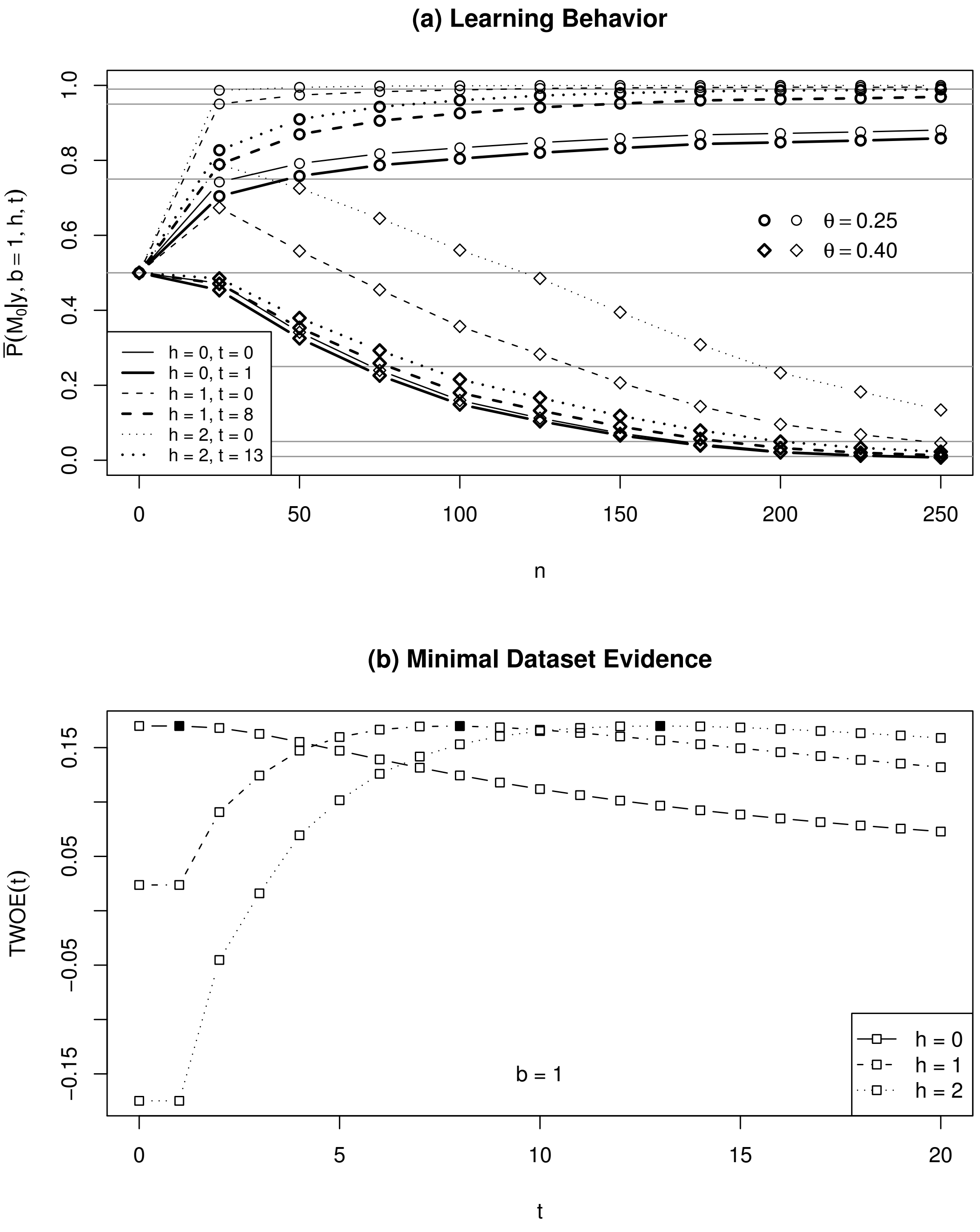

Figure 2: Learning behavior and minimal data set evidence for the Bernoulli example.

Next we discuss the choice of the hyperparameters h and t. The value of h determines the asymptotic behavior of $BF_{10}^{IM}(y^{(n)})$, as $n \to \infty$. The choice $h = 1$ is enough to change the convergence rate to $\mathcal{M}_0$ from sub-linear to super-linear (when $\mathcal{M}_0$ holds). Even though the convergence rate to $\mathcal{M}_1$ (when $\mathcal{M}_1$ holds) is unchanged, we have seen that increasing h induces a delay in learning $\mathcal{M}_1$. Thus, we recommend letting $h = 1$ by default, and trying $h = 2$ for sensitivity purposes. We discuss the choice of t in a separate subsection.

3.1 Choosing the training sample size

Recall that the goal of the intrinsic procedure is to pull mass toward the null subspace in the prior under $\mathcal{M}_1$. There is clearly a 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.

Fix $\theta_0 = 1/2$; this represents the worst scenario in terms of the information content of a single observation. In this case, the minimal sample size capable of providing evidence both in favor of the null and of the larger model is $n = 2$. Then, the data values $y = 0$ and $y = 2$ are in favor of $\mathcal{M}_1$, while the value $y = 1$ supports the null model. Consider the *weight of evidence* against the null using an intrinsic moment prior, i.e., $WOE_y(t) = \log BF_{10}^{IM}(y|b, h, t)$, where we focus on the dependence on t for a given choice of b, h (and given data y). Clearly, for symmetry, $WOE_0(t) = WOE_2(t)$.

Broadly speaking, the intrinsic approach rests on the following considerations: i) $WOE_1(t = 0)$ is too small; ii) $WOE_0(t = 0)$ and $WOE_2(t = 0)$ can be safely reduced without much harm. Point i) stems from the consideration that, when the data support the null model ($y = 1$ in our setup), and the prior is the default one ($t = 0$), the evidence in favor of $\mathcal{M}_1$ is too low for small sample sizes, because the default prior is too diffuse. On the other hand, as already remarked elsewhere, when the data clearly do not support $\mathcal{M}_0$ there will be enough evidence in favor of $\mathcal{M}_1$ for all reasonable (not overly diffuse) priors. Now, as t increases, so does $WOE_1(t)$, while $WOE_0(t)$ and $WOE_2(t)$ decrease, and point ii) comes into play.

As t diverges, the prior under $\mathcal{M}_1$ will progressively concentrate around the region defined by $\mathcal{M}_0$, so that the marginal data distributions under the two models will eventually coincide, and $WOE_y(t)$ will converge to zero, whatever the data y. What is then a natural minimal threshold for t to consider? 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 certainly 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 $t^* = \text{argmax}_t\, TWOE(t)$. The value t^* represents the *minimal* training sample size we should take into consideration when implementing the intrinsic procedure. Notice that this definition of minimal training sample size is *not* the usual one, which is adopted in the context of intrinsic priors or expected posterior priors (i.e., the smallest sample size such that the posterior is proper for all data outcomes). Also notice that, in practice, we need to check that t^* be well-defined. Then, we will probably be willing to let t vary over$\{t^*, \ldots, t^* + n\}$ for a sensitivity analysis.

We remark that the above strategy to find t^* in an intrinsic procedure is general, at least for discrete data models. In particular, the above strategy can be used to determine the minimal training sample size 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. It follows that, according to our criterion, when the starting prior is uniform, we could even dispense with the intrinsic procedure. 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

We consider as larger model the product of two binomial models

$$\mathcal{M}_1 : f_1(y_1, y_2|\theta_1, \theta_2) = \text{Bin}(y_1|n_1, \theta_1)\text{Bin}(y_2|n_1, \theta_2), \tag{11}$$

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

$$\mathcal{M}_0 : f_0(y_1, y_2|\theta) \;=\; \text{Bin}(y_1|n_1, \theta)\text{Bin}(y_2|n_1, \theta). \tag{12}$$

A default prior for θ under $\mathcal{M}_0$ is $p_0^D(\theta|b) = \text{Beta}(\theta|b, b)$, where $b = 1$ (or $b = 1/2$), while under $\mathcal{M}_1$ a default prior for (θ_1, θ_2) is $p_1^D(\theta_1, \theta_2|b) = \text{Beta}(\theta_1|b, b)\text{Beta}(\theta_2|b, b)$; in principle we could use different values of b for the two models, but we feel that little is lost by keeping our analysis simpler. For later purposes it is expedient to set the notation for a more general conjugate prior under $\mathcal{M}_1$, which we write as $p_1^C(\theta_1, \theta_2|a) = \text{Beta}(\theta_1|a_{11}, a_{12})\text{Beta}(\theta_2|a_{21}, a_{22})$, where $a = [[a_{jk}]_{k=1,2}]_{j=1,2}$ is a matrix of strictly positive real numbers. Then, we consider the conjugate moment prior

$$p_1^{CM}(\theta_1, \theta_2|a, h) = \frac{(\theta_1 - \theta_2)^{2h}}{K(a, h)}\text{Beta}(\theta_1|a_{11}, a_{12})\text{Beta}(\theta_2|a_{21}, a_{22}), \tag{13}$$

where

$$K(a, h) = \sum_{j=0}^{2h} \binom{2h}{j} (-1)^j \frac{B(a_{11} + j, a_{12})}{B(a_{11}, a_{12})} \frac{B(a_{21} + 2h - j, a_{22})}{B(a_{21}, a_{22})}. \tag{14}$$

As usual $h = 0$ returns $p_1^C(\theta_1, \theta_2|a)$. The default moment prior $p_1^{DM}(\theta_1, \theta_2|b, h)$ is obtained by letting $a_{11} = a_{12} = a_{21} = a_{22} = b$.

Consider now the intrinsic approach applied to p_1^{DM}. A natural requirement for an objective analysis is that the resulting joint prior for (θ_1, θ_2) be symmetric. It can be checked that, for the purely intrinsic case ($h = 0$), this happens even if the training sample sizes in the two groups, t_1 and t_2, are different (resulting from the fact that we use a single value of b for p_0^{DM} and p_1^{DM}). On the other hand, for the non-local case ($h > 0$), symmetry is only guaranteed if $t_1 = t_2 = t$ (the balanced case). We thus define the *intrinsic moment* prior of order h with training sample size t as

$$p_1^{IM}(\theta_1, \theta_2|b, h, t) = \sum_{x_1=0}^{t} \sum_{x_2=0}^{t} p_1^{DM}(\theta_1, \theta_2|x_1, x_2, b, h) m_0^D(x_1, x_2|b), \tag{15}$$

where

$$m_0^D(x_1, x_2|b) = \binom{t}{x_1}\binom{t}{x_2} \frac{B(b + x_1 + x_2, b + 2t - x_1 - x_2)}{B(b, b)}, \tag{16}$$

and $h = 0$ returns the standard intrinsic prior $p_1^I(\theta_1, \theta_2|b, t)$ with balanced training samples of size t. The default moment posterior in (15) can be computed as

$$p_1^{DM}(\theta_1, \theta_2|x_1, x_2, b, h) = p_1^{CM}(\theta_1, \theta_2|a_x^\star, h), \tag{17}$$

where $(a_x^\star)_{11} = b + x_1$, $(a_x^\star)_{12} = b + t - x_1$, $(a_x^\star)_{21} = b + x_2$, and $(a_x^\star)_{22} = b + t - x_2$.

Recall that the Bayes factor against $\mathcal{M}_0$ using an intrinsic moment prior under $\mathcal{M}_1$ is given by

$$BF_{10}^{IM}(y_1, y_2|b, h, t) = \sum_{x_1=0}^{t} \sum_{x_2=0}^{t} BF_{10}^{CM}(y_1, y_2|b, a_x^\star, h) m_0^D(x_1, x_2|b), \tag{18}$$

where $BF_{10}^{CM}(y_1, y_2|b, a_x^\star, h)$ is the Bayes factor based on the right hand side of (17). Similarly to the Bernoulli case, we can write

$$BF_{10}^{CM}(y_1, y_2|b, a, h) = \frac{K(a_y^\star, h)}{K(a, h)} BF_{10}^{C}(y_1, y_2|b, a), \tag{19}$$

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

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

and it follows that the Bayes factor against $\mathcal{M}_0$, using a conjugate moment prior under $\mathcal{M}_1$, can be written as

$$BF_{10}^{C}(y_1, y_2|b, a) = \frac{B(b, b) B(a_{11} + y_1, a_{12} + n_1 - y_1) B(a_{21} + y_2, a_{22} + n_2 - y_2)}{B(a_{11}, a_{12}) B(a_{21}, a_{22}) B(b + y_1 + y_2, b + n_1 + n_2 - y_1 - y_2)}. \tag{20}$$

Using (20) in (19) and plugging the latter into (18) 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. As in the Bernoulli case, we recommend choosing $b = 1$, which provides a uniform marginal distribution of $y_1 + y_2$ under $\mathcal{M}_0$ and of (y_1, y_2) under $\mathcal{M}_1$, and $h = 1$, which is enough to change the asymptotic behavior of the Bayes factor under the null from sub-linear to super-linear. As for the choice of t, we follow the general procedure outlined in the Bernoulli example, with suitable specific modifications to deal with the present case.

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, namely $(y_1 = 0, y_2 = 1)$ and $(y_1 = 1, y_2 = 0)$, are supportive for $\mathcal{M}_1$. We repeat the argument in subsection 3.1 and take $t^* = \text{argmax}_t \, TWOE(t)$ as the *minimal* training sample size, where $TWOE(t) = \sum_y WOE_y(t)$ and $WOE_y(t) = \log BF_{10}^{IM}(y_1, y_2|b, h, t)$.

Figure 3(a) plots $TWOE(t)$ for $h = 0, 1, 2$. As for the Bernoulli case, t^* is well-defined and when $h = 0$ (standard intrinsic prior) we get $t^* = 0$. Hence, we would recommend a sensitivity analysis with $t \in \{0, \min\{n_1, n_2\}\}$, say, in line with the analysis carried out by Casella and Moreno (2009, Table 2) on a collection of 2×2 tables, which we also examine later in this paper (section 5). On the other hand, when the starting prior is the default moment prior with $h = 1$ we find $t^* = 6$, while for $h = 2$ we get $t^* = 11$. Thus, as for 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, the comparison of these 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 valutes 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.

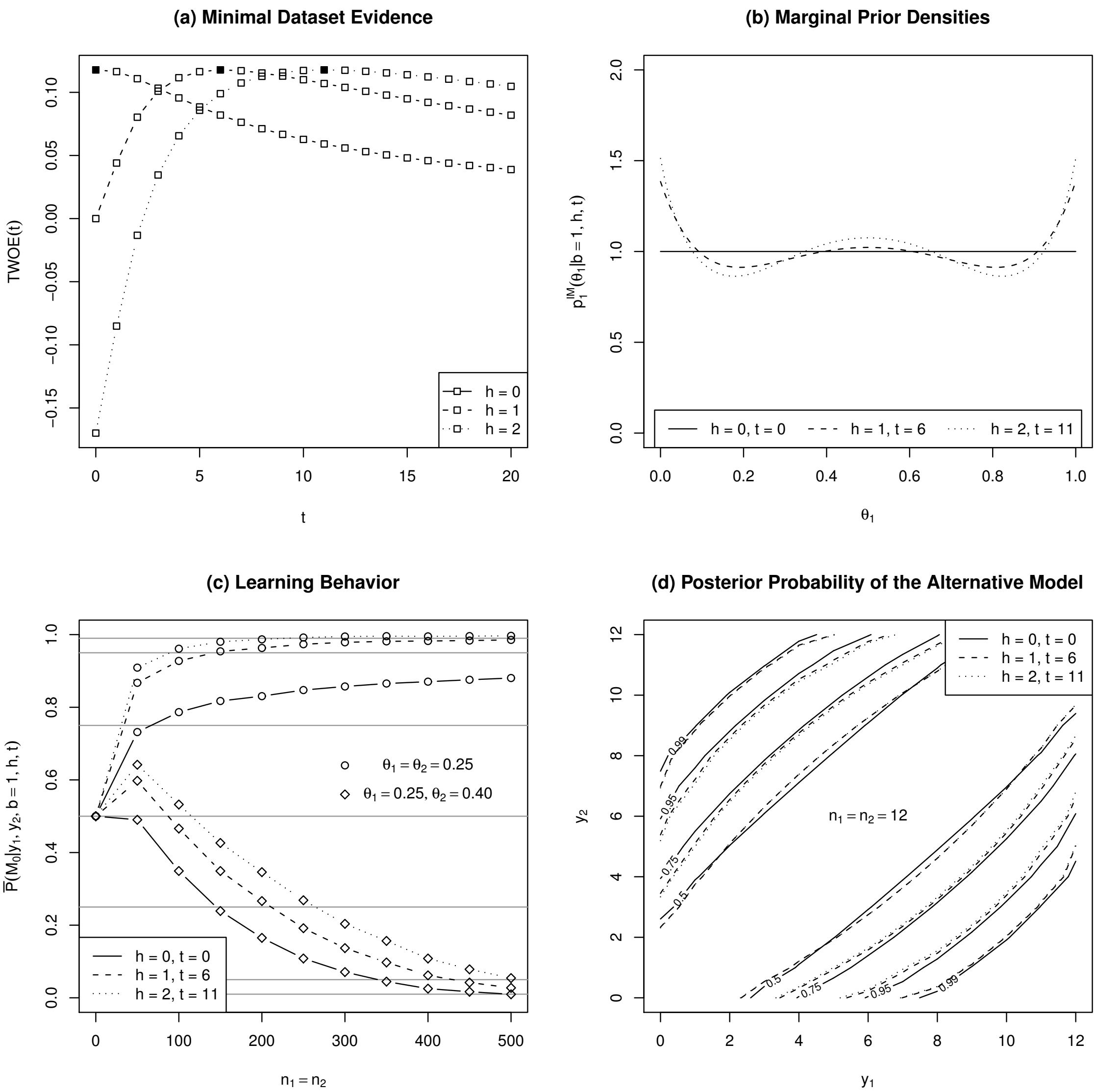

Figure 3: Characteristics of intrinsic moment priors for comparing two proportions.

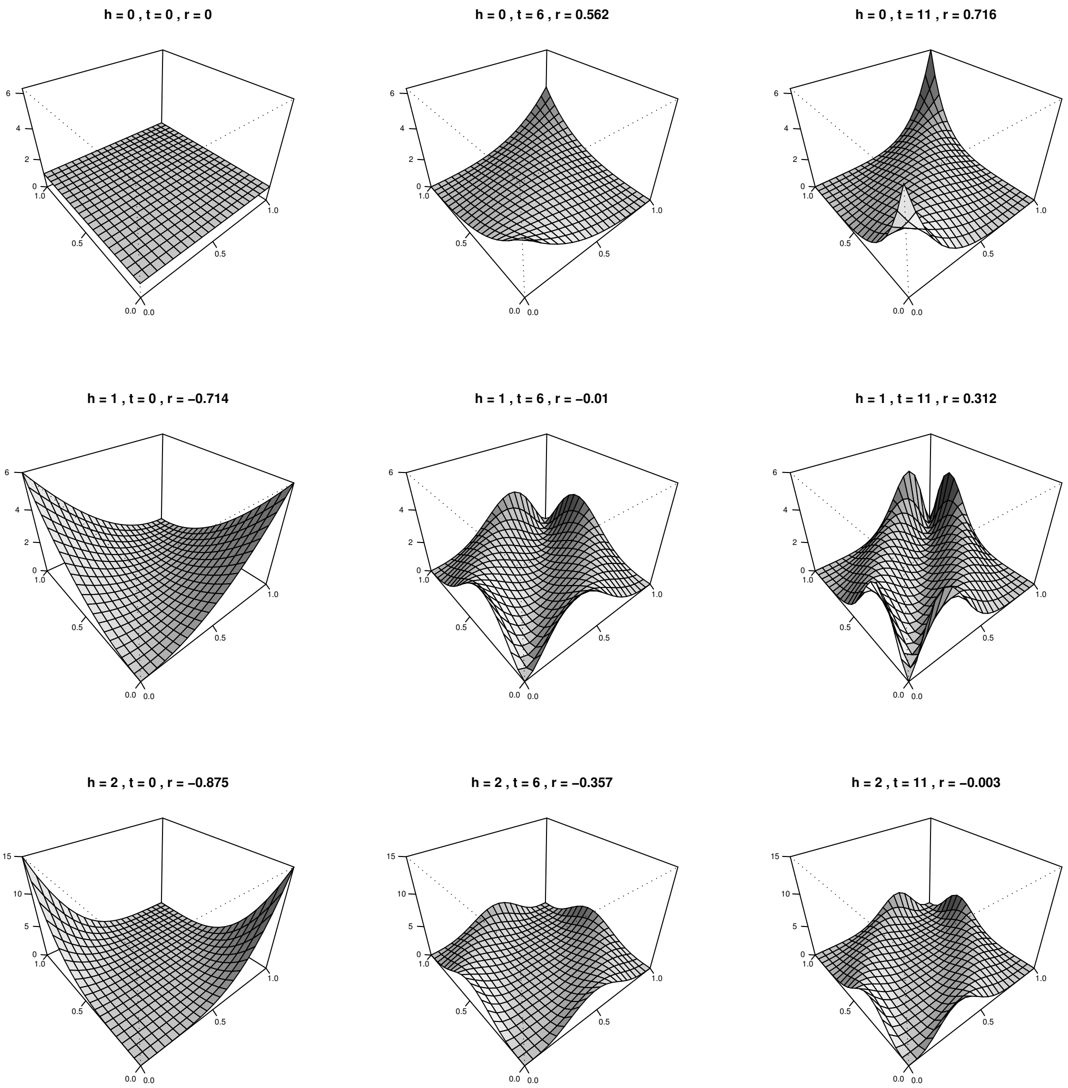

Figure 4: Intrinsic moment priors for comparing two proportions.

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 0.5, but the two intrinsic moment priors with $h > 0$ tend to moderately favor the outer values of the interval $(0, 1)$.

Figure 3(c) reports the average posterior probability of the null model (computed on 1000 simulated data sets of increasing size) letting first $\theta_1 = \theta_2 = 0.25$ and then $\theta_1 = 0.25, \theta_2 = 0.4$ (an instance of the alternative model). The learning behavior is quite different under the three priors. As for the Bernoulli example, when the data are generated under the null model a much quicker correct response is provided by the non-local intrinsic moment priors: for sample sizes up to 500 the average posterior probability of $\mathcal{M}_0$ under the default prior does not cross the 0.9 threshold, whereas under the non-local intrinsic moment priors it reaches the 95% threshold by the time 300 observations have been collected. On the other hand, switching from $h = 0$ to $h > 0$, the learning behavior under the alternative model is compromised in the short run, but not in the long run: there is an initial increase in the average posterior probability of the null model that takes about 100 observations to be neutralized, then the delay in learning stabilizes and by the time 500 observations have been collected strong evidence is achieved. These results suggest that the trade off in favor of $\mathcal{M}_1$ can be pushed further, when the intrinsic procedure is applied, and this provides motivation for a sensitivity analysis with $t > t^*$.

Figure 3(d) illustrates the small sample behaviour of intrinsic moment priors, by reporting the contour lines in the (y_1, y_2)-plane ($n_1 = n_2 = 12$) for selected thresholds of the posterior probability of $\mathcal{M}_1$. One can see, visually, a good degree of agreement among all three priors. There is also a clear indication that the higher thresholds, such as 90% and 95%, are reached for pairs (y_1, y_2) closer to the $y_1 = y_2$ line under the non-local intrinsic moment 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 Application

In this section 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 analyze the data using the intrinsic moment priors of section 4, letting $b = 1$ 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.

5.1 Sensitivity analysis

We let t vary from t^* to $t^* + \min\{n_1, n_2\}$ both for $h = 0$ (standard local prior) and $h = 1$ (recommended non-local prior), where $t^* = 0$ for $h = 0$ and $t^* = 6$ for $h = 1$ (minimal training sample size as discussed in subsection 4.1), while n_1 and n_2 are the trial sample sizes for Treatment and Control. For each of the above pairs (h, t), and for all 41 tables in the dataset, we evaluate the posterior probability of the null model. It turns out that the latter is quite insensitive to any further increase in t (beyond $t^* + \min\{n_1, n_2\}$). 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 proportions): 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 $t = t^* + \min\{n_1, n_2\}$ are marked with circles and triangles, respectively, so that in practice (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.

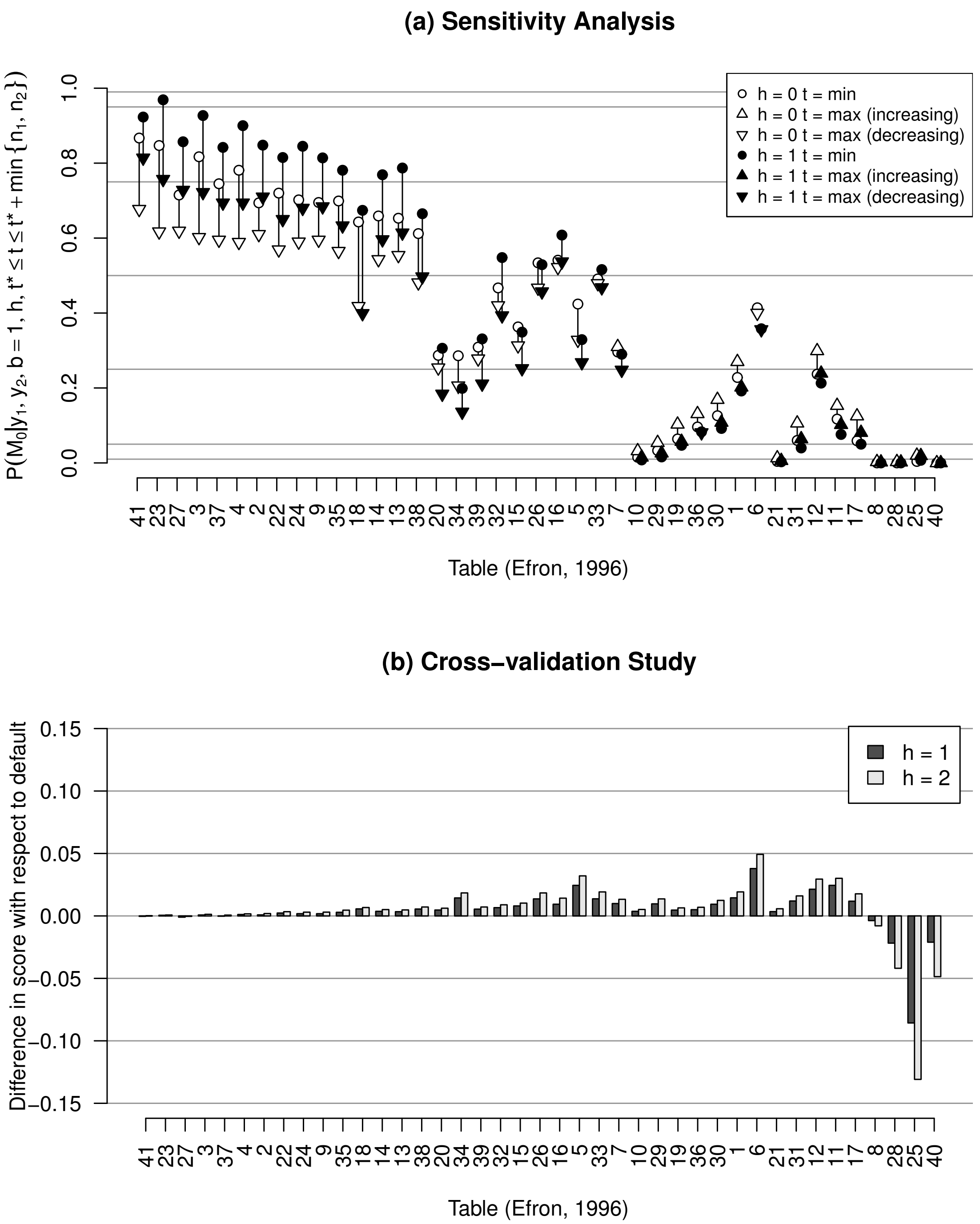

Figure 5: Results of the sensitivity analysis and cross-validation study: each number on the horizontal axis identifies a table.

Consider first the left-hand tables. Except for table 18 (and possibly table 38) the posterior probability of $\mathcal{M}_0$ ranges 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 arrow) produces values for the posterior probability of $\mathcal{M}_0$ higher than under the standard intrinsic prior (white arrow): this is only to be expected, because of the local *versus* non-local nature of these priors. 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 $\mathcal{M}_0$. Save for table 18 (and possibly 38) a robust conclusion can be reached in favor of the equality of proportions between the two groups. Next consider the tables in the center. Here the lengths of the intervals are shorter than before, with the majority of tables exhibiting a posterior probability of the null below the 0.5 threshold, and the remaining ones hovering over it. Again all arrows point downwards, indicating that the intrinsic procedure is working in favor of the alternative, although to a much lesser extent than for the left-hand tables. This makes sense because $\mathcal{M}_0$ is less supported in this group of tables, and hence the amount of evidence that can be transferred to $\mathcal{M}_1$ is limited. The conclusion against the equality of the two proportions is robust for the majority of the tables in this group, but the analysis is inconclusive for tables 32, 26 and 33. Finally, the pattern of the right-hand tables indicates a low support for the null, with the possible exception of table 6. Most of the arrows point upwards, but all ranges are very short and on some occasions negligible: this is the action of the intrinsic procedure in favor of $\mathcal{M}_0$, because the data do not support the null model.

5.2 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). 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 an occurrence in group 1. We exclude this case from the dataset and compute $\hat{\theta}_1^{(1)}$ 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 $(y_1 - 1, n_1 - y_1, y_2, n_2 - y_2)$; similarly, for an occurrence in group 2, we compute $\hat{\theta}_2^{(1)}$ upon interchanging subscript 1 and 2 above. On the other hand, to predict a nonoccurrence in group 1, we let $\hat{\theta}_1^{(0)}$ be 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 $\hat{\theta}_2^{(0)}$ to predict a nonoccurrence in group 2 requires interchanging subscript 1 and 2. In the spirit of cross-validation, we repeat the analysis for each case and compute the overall mean score

$$S = \frac{y_1 \log \hat{\theta}_1^{(1)} + (n_1 - y_1)\log(1 - \hat{\theta}_1^{(0)}) + y_2 \log \hat{\theta}_2^{(1)} + (n_2 - y_2)\log(1 - \hat{\theta}_2^{(0)})}{n_1 + n_2}. \qquad (21)$$

Now let S_h be the score associated to 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 uniform 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 proportions. Essentially for all tables, but with the notable exception of the last four, the non-local intrinsic moment priors perform better than the standard intrinsic prior, with differences in score ranging from −0.1% to 3.8% (median improvement 0.5%) when $h = 1$ and from 0.0% to 4.9% (median improvement 0.7%) when $h = 2$. On the other hand, for the last four tables, which are clearly against the null, the performance of the 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 -0.4% down to -8.6%, when $h = 1$, and from -0.8% down to -13.1%, 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 Discussion

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

A robustness analysis is naturally embedded in our approach, by letting the training sample size vary over a grid of values. The notion of minimal training sample size is more delicate to handle in our case than in the case of the standard intrinsic approach. To this aim, we devised the notion of total weight of evidence as a natural currency to trade evidence stakes (on the log scale). While this notion worked fine in our problems, its broad applicability still remains an open issue and should be carefully evaluated in each specific case. In particular, it would be interesting to see how far its scope could be extended beyond discrete data models.

The general methodological framework developed in this paper was tried out on a substantive statistical issue, namely testing the equality of two independent proportions. This resulted in a novel objective Bayesian test for this problem, which leads to sharper conclusions, even for moderately large samples, when the two population proportions are actually equal. An extension of our methods to testing independence in general contingency tables under a variety of sampling schemes, as in Casella and Moreno (2009), would constitute a natural and useful development.

References

Albert, J. H. (1990), "A Bayesian Test for a Two-way Contingency Table Using Independence Priors," *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 18, 347–363.

Albert, J. H. and Gupta, A. K. (1982), "Mixtures of Dirichlet Distributions and Estimation in Contingency Tables," *The Annals of Statistics*, 10, 1261–1268.

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.

Casella, G., Girón, F. J., Martínez, M. L., and Moreno, E. (2009), "Consistency of Bayesian Procedures for Variable Selection," *The Annals of Statistics*, 37, 1207–1228.

Casella, G. and Moreno, E. (2005), "Intrinsic Meta-analysis of Contingency Tables," *Statistics in Medicine*, 24, 583–604.

— (2006), "Objective Bayesian Variable Selection," *Journal of the American Statistical Association*, 101, 157–167.

— (2009), "Assessing Robustness of Intrinsic Tests of Independence in Two-Way Contingency Tables," *Journal of the American Statistical Association*, 104, 1261–1271.

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.

Dawid, A. P. (1999), "The Trouble with Bayes Factors," Research Report 202, University College London Department of Statistical Science, `http://www.ucl.ac.uk/Stats/research/reports/abs99.html#202`.

Efron, B. (1996), "Empirical Bayes Methods for Combining Likelihoods," *Journal of the American Statistical Association*, 91, 538–550, with discussion: 551–565.

Girón, F. J., Moreno, E., and Casella, G. (2007), "Objective Bayesian Analysis of Multiple Changepoints for Linear Models," in *Bayesian Statistics 8*, eds. Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M., Oxford University Press, pp. 227–252.

Howard, J. V. (1998), "The 2×2 Table: a Discussion from a Bayesian Viewpoint," *Statistical Science*, 13, 351–367.

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.

Kass, R. E. and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Moreno, E. (1997), "Bayes Factors for Intrinsic and Fractional Priors in Nested Models. Bayesian Robustness," in *L_1-Statistical Procedures and Related Topics*, ed. Dodge, Y., Institute of Mathematical Statistics, pp. 257–270.

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.

Morris, C. M. (1987), "Discussion of Berger/Sellke and Casella/Berger," *Journal of the American Statistical Association*, 82, 106–111.

Perez, 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*, Springer-Verlag, 2nd ed.