Journal of Mathematical Psychology 80 (2017) 33–39

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

A theoretical note on the prior information criterion

Sara Steegen [a], Woojae Kim [b], Wiebe Pestman [a], Francis Tuerlinckx [a], Wolf Vanpaemel [a,*]

[a] KU Leuven, Leuven, Belgium
[b] Howard University, Washington, DC, USA

HIGHLIGHTS

- We consider the prior information criterion and compare it to the Bayes factor.
- We use binomial models as an application example to test inequality and equality constraints.
- In contrast to the Bayes factor, the prior information criterion can yield inconsistent selection results.
- We evaluate analytic forms and present a formal relationship between the two methods.

ARTICLE INFO

Article history:
Received 9 February 2017
Received in revised form 5 June 2017
Available online 12 August 2017

Keywords:
Prior information criterion
Bayes factor
Bayesian model selection
Binomial model

ABSTRACT

We consider the recently proposed *prior information criterion* for statistical model selection (PIC; van de Schoot et al. 2012). Using simple binomial models as an example, we demonstrate that the PIC can produce puzzling outcomes. When employed to test various forms of inequality and equality constraints, the PIC can yield inconsistent selection results, in that it fails to select the correct, data-generating model even when the underlying truth lies strictly in that model, and not in the alternative model. Moreover, in certain cases, such inconsistency arises for all sample sizes, meaning that it is not merely an asymptotic property. By contrast, when applied across the same testing scenarios, the Bayes factor provides consistent model selection. We explain why the PIC exhibits inconsistent model selection by examining its analytic forms for binomial models in comparison to those of the Bayes factor. We extend the same account to exponential families, and provide an insight into general cases in which the PIC bears a relationship to the Bayes factor.

© 2017 Elsevier Inc. All rights reserved.

In psychology, as well as in other scientific fields, statistical models are used to describe the underlying processes of probabilistic phenomena under study and thereby explain the regularities behind observed data. In developing such models, it is often the case that different plausible accounts are proposed. Toward providing an objective criterion for testing competing models, the development of model selection methods has been an important topic of research (e.g., Claeskens & Hjort, 2008). With this goal in mind, van de Schoot, Hoijtink, Romeijn, and Brugman (2012) proposed the *prior information criterion* (PIC) as a Bayesian method for testing models under (in)equality constraints (e.g., whether a certain parameter is greater than, less than, or equal to a fixed value).

For a model with parameter θ, likelihood $f(y|\theta)$ and prior $p(\theta)$, the PIC is defined as

$$\begin{aligned} \text{PIC} &= E_{p(\theta)}\left[-2\log f(y|\theta)\right] \\ &= -2\int \log\left(f(y|\theta)\right) p(\theta)\,\mathrm{d}\theta. \end{aligned} \tag{1}$$

As $\log f(y|\theta)$ is often used to indicate model-data fit, the PIC reflects a lack of fit. Selection between two models is based on the difference in their PIC values, $\Delta\text{PIC} = \text{PIC}_1 - \text{PIC}_2$:

$$\begin{aligned} \Delta\text{PIC} &= -2\int \log\left(f_1(y|\theta_1)\right) p_1(\theta_1)\,\mathrm{d}\theta_1 \\ &\quad + 2\int \log\left(f_2(y|\theta_2)\right) p_2(\theta_2)\,\mathrm{d}\theta_2. \end{aligned} \tag{2}$$

A negative value of ΔPIC indicates a preference for Model 1 over Model 2, whereas a positive value indicates a preference for Model 2 over Model 1.

As noted by van de Schoot et al. (2012), ΔPIC is closely related to the Bayes factor (BF; Jeffreys, 1961). The Bayes factor is the ratio

* Correspondence to: Faculty of Psychology and Educational Sciences, Tiensestraat 102, bus 3713, 3000 Leuven, Belgium. Fax: +32 16325993.
E-mail address: wolf.vanpaemel@kuleuven.be (W. Vanpaemel).

http://dx.doi.org/10.1016/j.jmp.2017.06.002
0022-2496/© 2017 Elsevier Inc. All rights reserved.

of two marginal likelihoods (ML), with the ML defined as

$$\begin{aligned} \text{ML} &= E_{p(\theta)}[f(y|\theta)] \\ &= \int f(y|\theta)p(\theta)\,\mathrm{d}\theta. \end{aligned} \tag{3}$$

The marginal likelihood integrates over the entire prior distribution, and thus reflects the average fit of a model to the data, weighted by the prior beliefs about the model parameter θ. In this paper, we will express the Bayes factor $\text{BF} = \frac{\text{ML}_1}{\text{ML}_2}$ in the following form: $-2\log\text{BF} = -2\log\text{ML}_1 + 2\log\text{ML}_2$. This way, we obtain a model selection criterion on the same scale as the well-known likelihood-ratio test statistic or the deviance (Kass & Raftery, 1995):

$$\begin{aligned} -2\log\text{BF} = &-2\log\int f_1(y|\theta_1)p_1(\theta_1)\,\mathrm{d}\theta_1 \\ &+2\log\int f_2(y|\theta_2)p_2(\theta_2)\,\mathrm{d}\theta_2. \end{aligned} \tag{4}$$

As with ΔPIC, a negative value of $-2\log$BF indicates that Model 1 is more likely than Model 2, whereas a positive value indicates the opposite.

A clear commonality between the PIC and the ML is their sensitivity to the prior. While this sensitivity is sometimes seen as a drawback, it can also be considered an advantage. One major form of Bayesian hypothesis testing involves specifying different priors for a given likelihood, each corresponding to different hypotheses, and comparing the resulting models given observed data. In this case, sensitivity to the prior is an advantage rather than a drawback (e.g., Vanpaemel, 2010). For example, using the PIC or the Bayes factor, one can check whether a regression coefficient is greater than a certain value by testing a model that incorporates such a hypothesis in the form of a constrained prior.

Besides their similar forms, ΔPIC and $-2\log$BF are two distinct model selection criteria. Their technical difference is apparent from the location of the logarithm operator: The PIC places the logarithm inside the integral (Eq. (2)) whereas the BF places it outside the integral (Eq. (4)). In this paper, we illustrate applications in which this difference leads to very different model selection outcomes of the two criteria. In particular, we report that the application of the PIC produces unexpected and puzzling results, which can be explained based on its analytic forms. Further, we present a formal relationship between the two criteria for general cases and provide insight into their disparate behavior.

1. Application to binomial models

We investigate the behavior of the PIC and the Bayes factor in the context of the very basic, but widely used, binomial model with a single parameter θ representing the probability of success. In creating multiple scenarios of hypothesis testing, both inequality and equality constraints on the parameter are considered. Although the original motivation behind the PIC's development was to propose a method for testing inequality constrained hypotheses, van de Schoot et al. (2012) themselves applied the PIC to test an equality constrained hypothesis as well (e.g., Real-Life Example 1, pp. 14–15, van de Schoot et al., 2012). Moreover, we see no rationale in the derivation of the PIC that fundamentally precludes its application in testing equality constraints.

Consider the number of successes $y = 0, 1, \ldots, n$ with the binomial probability mass function,

$$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \tag{5}$$

where $0 < \theta < 1$ is the probability of success. We focus on four models representing different hypotheses about θ, which are specified by different priors for θ: an unconstrained model (M_{01} : $0 < \theta < 1$), a one-sided inequality constrained model (M_{0u} : $0 < \theta < u$), a two-sided inequality constrained model (M_{lu} : $l < \theta < u$), and an equality constrained model (M_c : $\theta = c$). The unconstrained and inequality constrained models assume uniform distributions whose supports are defined by their hypothesized intervals. The equality constrained model places a point mass at c (see Fig. 1).

Given these models, our analysis of the behavior of the PIC and the Bayes factor considers three scenarios in which each of the constrained models is tested against the unconstrained model (i.e., M_{0u} vs. M_{01}, M_{lu} vs. M_{01}, and M_c vs. M_{01}). To conduct these tests, we derived analytic-form expressions of the PIC and the ML for each model (see Supplementary Materials, Appendix A). These are summarized in Table 1 and illustrated in Fig. 2 in the forms of ΔPIC and $-2\log$BF for each scenario. In particular, Fig. 2 displays the behavior of ΔPIC and $-2\log$BF applied to the test of the constrained models against the unconstrained model for every possible sample proportion y/n when $n = 100$. For each type of the constrained hypotheses (each row), three examples of constraints (each column) are considered. For each method, a positive value indicates a preference for the constrained model over the unconstrained model, and a negative value does the opposite, as denoted on the leftmost vertical axis.

1.1. Testing one-sided inequality constrained hypotheses

The top row of Fig. 2 shows the values of ΔPIC and $-2\log$BF as a function of observed proportions when testing a one-sided inequality constrained hypothesis (M_{0u} : $0 < \theta < u$) against the unconstrained model (M_{01} : $0 < \theta < 1$). As seen in the figure, in this comparison scenario, both criteria decrease with the increasing sample proportion y/n, which makes good sense. Both methods select M_{01} if high values of y/n are observed and select M_{0u} if low values of y/n are observed. Moreover, consistent with what would be expected, both methods prefer M_{01} more often if M_{0u} places a more strict constraint on θ (i.e., a smaller value of u).

Despite these similarities, there are two clear differences between the two methods. First, the trend of ΔPIC over y/n is linear whereas that of $-2\log$BF is curved. The second difference concerns the point on y/n where their values cross zero, or the *decision bound*, at which neither of the two models is clearly supported by the data. Consider the case of ΔPIC first. Analytic solutions for ΔPIC's decision bounds for each model comparison are listed in Table 2. Inspection of the table reveals that the decision bound of ΔPIC on y/n is constant and not affected by the sample size n. This fact provides a perspective from which to better view the behavior of ΔPIC seen in Fig. 2. When M_{0u} with $u = .3$ is compared against M_{01} (upper left panel in Fig. 2), the decision bound of ΔPIC is approximately .41, meaning that when the observed proportion y/n is less than .41, the constrained model is favored. According to the closed-form expressions in Table 2 (though not visible in Fig. 2), this will be the case no matter what the sample size n is.

The fact that the decision bound of ΔPIC is insensitive to n has a significant implication for its behavior that is distinguished from that of $-2\log$BF. Suppose that the underlying process is a binomial model whose success probability θ_0 lies above the upper bound u of the constraint, but below the decision bound (e.g., $\theta_0 = .35$ when $u = .3$). Then, it is reasonable to expect that, once a sufficient amount of data is collected, the unconstrained model (M_{01} : $0 < \theta < 1$) should almost surely be identified against the incorrect constrained model (M_{0u} : $0 < \theta < u$). The selection outcome of ΔPIC contradicts this intuition: As n increases, by the central limit theorem, the observed proportion y/n concentrates at $\theta_0 = .35$ and since this is below the decision bound (i.e., .41), ΔPIC will incorrectly select the constrained model. Thus, even when the

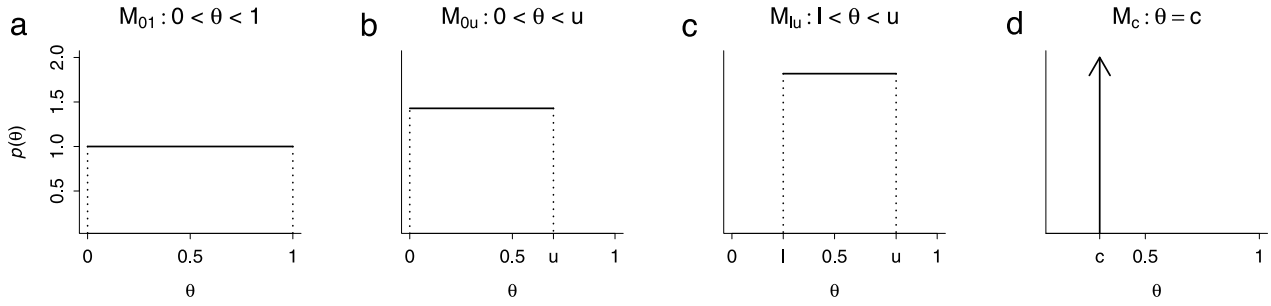


Fig. 1. Prior distributions for θ, specifying the unconstrained model (a), one-sided inequality constrained model (b), two-sided inequality constrained model (c), and equality constrained model (d).

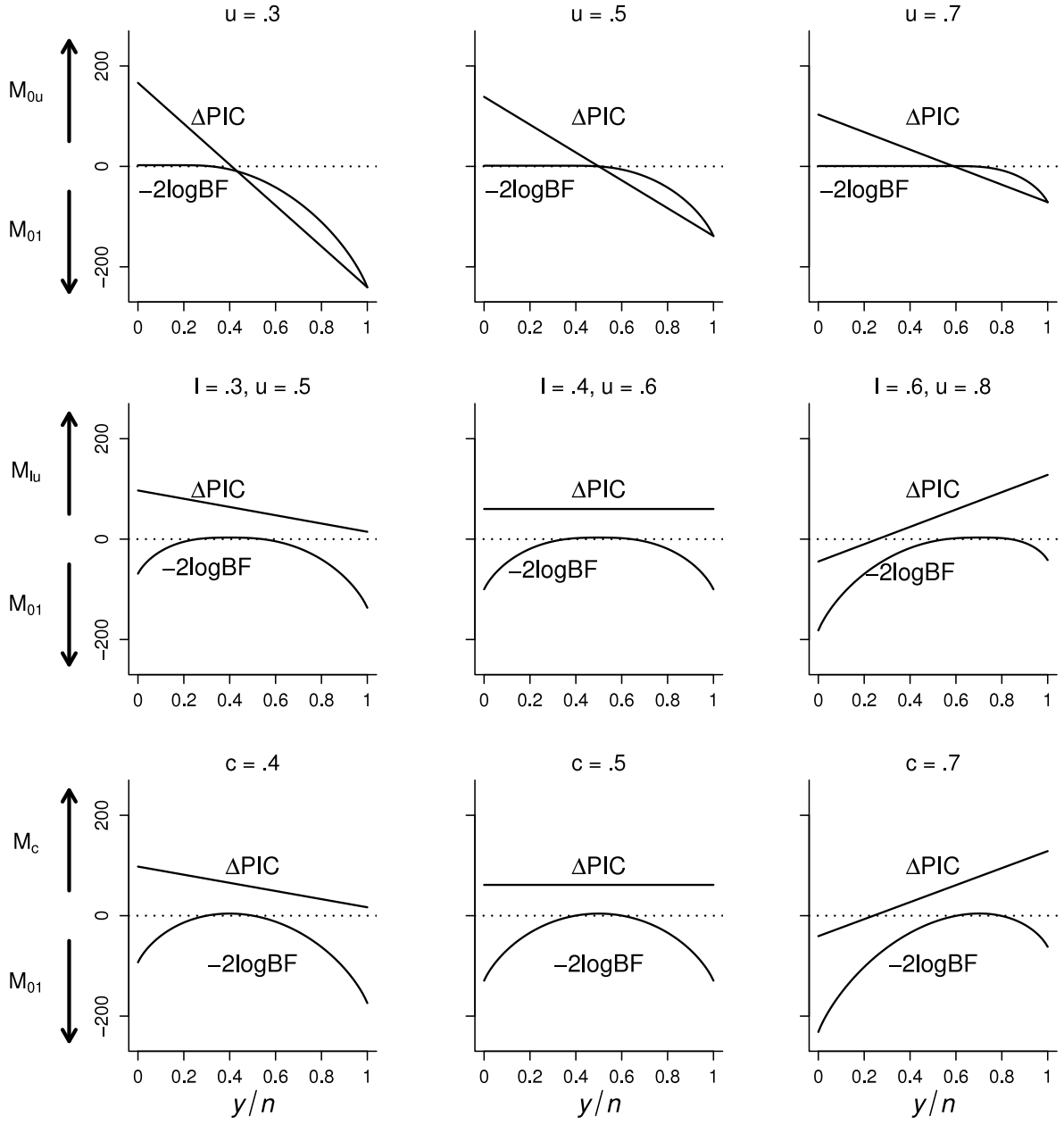


Fig. 2. Pairwise comparison of binomial models under different hypotheses based on the PIC and BF when $n = 100$. For both ΔPIC and -2 log BF, a positive value indicates a preference for the constrained model and a negative value indicates a preference for the unconstrained model. Top row: Test of the one-sided inequality constrained model ($M_{0u} : 0 < \theta < u$) against the unconstrained model ($M_{01} : 0 < \theta < 1$). Middle row: Test of the two-sided inequality constrained model ($M_{lu} : l < \theta < u$) against the unconstrained model ($M_{01} : 0 < \theta < 1$). Bottom row: Test of the equality constrained model ($M_c : \theta = c$) against the unconstrained model ($M_{01} : 0 < \theta < 1$).

Table 1
Analytic expressions of ΔPIC and $-2\log$BF applied to three pairwise comparisons of four binomial models with different priors: The unconstrained model (M_{01}) with a uniform prior on the interval [0, 1], the one-sided inequality constrained model (M_{0u}) with a uniform prior on the interval [0, u], the two-sided inequality constrained model (M_{lu}) with a uniform prior on the interval [l, u], and the equality constrained model (M_c) specified by a Dirac delta function at c. The derivations can be found in the Supplementary Materials (see Appendix A).

M2 vs. M1	ΔPIC
$M_{0u} : 0 < \theta < u$ vs. $M_{01} : 0 < \theta < 1$	$2y\big(\log(u) - \frac{u-1}{u}\log(1-u)\big) + 2n\big(\frac{u-1}{u}\log(1-u)\big)$
$M_{lu} : l < \theta < u$ vs. $M_{01} : 0 < \theta < 1$	$-2y\frac{-(1-l)\log(1-l) + u\log(u) + (1-u)\log(1-u) - l\log(l)}{u-l}$ $- 2n\frac{(1-l)(\log(1-l)-1) - (1-u)(\log(1-u)-1)}{u-l}$
$M_c : \theta = c$ vs. $M_{01} : 0 < \theta < 1$	$2y\left(\log\frac{c}{1-c}\right) + 2n\big(1+\log(1-c)\big)$

M2 vs. M1	$-2\log$BF
$M_{0u} : 0 < \theta < u$ vs. $M_{01} : 0 < \theta < 1$	$2\log\frac{I_u(y+1, n-y+1)}{u}$
$M_{lu} : l < \theta < u$ vs. $M_{01} : 0 < \theta < 1$	$2\log\frac{I_u(y+1, n-y+1) - I_l(y+1, n-y+1)}{u-l}$
$M_c : \theta = c$ vs. $M_{01} : 0 < \theta < 1$	$2y\left(\log\frac{c}{1-c}\right) - 2\log B(y+1, n-y+1) + 2n\big(\log(1-c)\big)$

Note: y is the observed number of successes, n is the number of trials, l, u and c are constants, $B(\cdot,\cdot)$ is the beta function, $B_u(\cdot,\cdot)$ is the incomplete beta function, and $I_u(\cdot,\cdot)$ is the regularized incomplete beta function.

Table 2
Decision bounds of ΔPIC for the three pairwise comparisons. At decision bounds, the analytic expressions of ΔPIC in Table 1 equal zero, indicating no model preference.

M2 vs. M1	Decision bound
$M_{0u} : 0 < \theta < u$ vs. $M_{01} : 0 < \theta < 1$	$-\frac{-(1-u)\log(1-u)}{(1-u)\log(1-u) + u\log(u)}$
$M_{lu} : l < \theta < u$ vs. $M_{01} : 0 < \theta < 1$	$-\frac{(1-l)\log(1-l) - (1-u)\log(1-u)}{-(1-l)\log(1-l) + (1-u)\log(1-u) + u\log(u) - l\log(l)}$
$M_c : \theta = c$ vs. $M_{01} : 0 < \theta < 1$	$-\frac{1+\log(1-c)}{\log\frac{c}{1-c}}$

truth lies *only* in one model and not in the alternative one, the PIC can prevent one from selecting the model containing the truth.[1]

By contrast, in the same situation, $-2\log$BF behaves as anticipated. This can be seen as follows. In the expression of $-2\log$BF for testing M_{0u} versus M_{01}, shown in Table 1, the numerator of the fraction inside the logarithm (i.e., $I_u(y+1, n-y+1)$) is in fact $p(\theta < u|y)$, the posterior probability of θ being less than u. Since, as per the large-sample property of Bayesian posteriors (Schervish, 1995), the posterior distribution of θ concentrates at the true proportion θ_0 as n increases, $p(\theta < u|y)$ converges to 1 when $\theta_0 < u$ and to zero when $\theta_0 > u$. Therefore, the selection based on $-2\log$BF will be consistent with the underlying process as data accumulate.

[1] One may wonder if the behavior of ΔPIC may be due to a particular specification of the prior distribution (i.e., a uniform distribution in the current example). The decision bound indeed depends on the prior. In fact, in the case of $u = .5$ under the uniform prior, the decision bound happens to be .5, which will prevent counterintuitive selection behavior. Even for another value of u, the decision bound can be made to avoid such illogical selection by using a different prior (our derivation of the PIC in the Supplementary Materials (see Appendix A) assumes a general class of beta priors). However, a proper Bayesian inference should not require one to confine models to a particular prior specification in order to achieve consistency in model selection. Normally, the effect of priors is overridden by a sufficient accumulation of evidence in data. This is precisely the behavior of Bayes factors, which has been proved for general cases (Doob, 1949; Schwartz, 1965), but, as shown in our analysis, is not exhibited by ΔPIC.

1.2. Testing two-sided inequality constrained hypotheses

The test of a two-sided inequality constrained hypothesis (i.e., $M_{lu} : l < \theta < u$ vs. $M_{01} : 0 < \theta < 1$) is shown in the middle row of Fig. 2. In this comparison scenario, the PIC and the BF exhibit more drastic differences in their selection behavior. Again, ΔPIC's linear versus $-2\log$BF's nonlinear relationships to the data y/n are apparent. In this case, however, a selection criterion's nonlinear response to the evidence in data is vital for its outcome to make sense: It is supposed to select M_{01} when the observed proportion is sufficiently far away from the hypothesized range of M_{lu} ($l < \theta < u$) by being either small or large, and it should favor M_{lu} when the observation is within or near the range.

However, ΔPIC's selection behavior belies this intuition. When testing $M_{lu} : .3 < \theta < .5$ (leftmost panel), ΔPIC always favors M_{lu} no matter what proportion is observed, even when it is very small or very large. In the case of testing $M_{lu} : .4 < \theta < .6$ (middle panel), the quantity itself of ΔPIC is completely insensitive to the data as it is reduced to a positive constant. The closed-form expression in Table 1 shows that this will always be the case whenever $u = 1 - l$. When $M_{lu} : .6 < \theta < .8$ is tested (rightmost panel), in which the hypothesized range shifts closer to an extreme proportion, ΔPIC finally allows for the possibility that M_{01} is selected, but only when very small proportions are observed. Still, ΔPIC does not accept any evidence for M_{01} when the proportion lies beyond the other side of the constraint. For

instance, ΔPIC prefers $M_{lu} : .6 < \theta < .8$ even with 100 successes out of 100 trials.

Note that, unlike in the previous comparison scenario involving one-sided constraints, we do not need to postulate a large-sample condition in order to see the PIC's counterintuitive selection behavior. For the test of two-sided inequality constraints, its illogical selection can occur for *any* data, regardless of the sample size. To see this, suppose that the data-generating process is a binomial model with θ_0 located outside the interval (l, u) so that we expect that with sufficient data M_{01} should almost surely be preferred over M_{lu}. For example, suppose that the underlying truth is $\theta_0 = .2$ or $\theta_0 = .7$ when $M_{lu} : .3 < \theta < .5$ is tested (e.g., consider the first example in the middle row of Fig. 2). In this situation, ΔPIC, as a linear function of y, has a decision bound beyond 1, which is the logical limit of y/n. Since Table 2 shows the decision bound is not affected by n, the PIC will never select M_{01}, no matter what data in any size sensibly support M_{01}.

More generally, it is the very nature of the constraint $l < \theta < u$, bounded on two sides, that prevents the PIC from handling such a hypothesis properly. Whenever an inequality constraint of this type is tested, simply because ΔPIC, as a linear function of data, cannot form two decision bounds, there always exists an underlying proportion θ_0 outside the interval (l, u) that makes it impossible for the PIC to favor M_{01} with any data of any amount. In some cases, the PIC will never select M_{01} for all values of the underlying truth θ_0 outside (l, u) no matter what data evidence M_{01} (e.g., examples in the leftmost and middle panels of the figure), and in other cases, data in support of M_{01} cannot arise if θ_0 is either very small or very large (e.g., the example in the rightmost panel).

In contrast, again, the BF performs as expected: $-2 \log$ BF responds to the data nonlinearly, favoring M_{01} when the sample proportion is far from the two-sided constraint (l, u), and M_{lu} when the observation is in or close to (l, u). This is reflected in its inverted U-shapes in all three examples of (l, u) in Fig. 2. Asymptotically, in the expression of $-2 \log$ BF for testing M_{lu} versus M_{01} in Table 1, the numerator inside the logarithm equals $p(l < \theta < u|y)$, which converges to 1 when $l < \theta_0 < u$ and to zero when $\theta_0 < l$ or $\theta_0 > u$ as n increases. Therefore, the decision based on $-2 \log$ BF is consistent.

1.3. Testing equality constrained hypotheses

The results from testing an equality constrained hypothesis (i.e., $M_c : \theta = c$ vs. $M_{01} : 0 < \theta < 1$) are shown in the bottom row of Fig. 2. The selection pattern of the two methods across the data y/n is qualitatively the same as in the previous scenario of testing two-sided inequality constrained hypotheses illustrated in the middle row of the figure. With the hypothesis $\theta = .4$ (leftmost panel), ΔPIC is above zero irrespective of data, thus always selecting M_c. When testing $\theta = .5$ (middle panel), ΔPIC becomes a positive constant, again always selecting M_c. Only when the hypothesized value for θ is away from .5 to some extent like $c = .7$ (rightmost panel), ΔPIC allows M_{01} to be selected for very small y/n (decision bound $\approx .24$), but not for large proportions.

Essentially, the properties of the two methods described previously still hold here: A situation exists in which the PIC will never favor M_{01}, no matter what data in any size sensibly support M_{01}, whereas the BF performs as expected. In testing the equality constraint $\theta = c$, a selection criterion is expected to prefer M_{01} when y/n is away from the hypothesized value c by being either small or large, and it should favor M_c when y/n is equal or close to c. ΔPIC cannot accomplish this by nature, because it is a linear function of data and cannot provide two decision bounds. In addition, as with earlier tests, the decision bound of ΔPIC for M_c versus M_{01} is constant for all sample sizes as shown in Table 2. Consequently, there always exists an underlying proportion $\theta_0 \neq c$ that makes it impossible for the PIC to select M_{01} given any data. In some cases (e.g., $c = .4$ or $c = .5$; i.e., leftmost and middle panels), the PIC will never choose M_{01} for all values of the underlying truth $\theta_0 \neq c$ no matter what data support M_{01}. In other cases, data evidencing M_{01} cannot arise if θ_0 is either very small or very large (e.g., in the rightmost panel).

The BF performs as a reasonable criterion should, selecting M_c only when the observed proportion is close to c, the hypothesized value of θ. Its selection is again consistent, which can be seen in the analytic form of $-2 \log$ BF for M_c versus M_{01} in Table 1. In this case, $-2 \log$ BF equals $2 \log p(\theta = c|y)$, twice the log posterior density function of θ evaluated at c, which goes to ∞ if $\theta_0 = c$, and to $-\infty$ if $\theta_0 \neq c$ as n increases, due to the asymptotic convergence of a posterior distribution.

2. Generalization

We have shown two peculiar properties of the PIC when applied to the comparison of a binomial model with an inequality or an equality constraint on its success probability against the encompassing, unconstrained model. First, ΔPIC is always a linear function of the observed proportion, and as a result, provides one-sided evidence, responding to only either small or large proportions in support of one of the models. Second, ΔPIC is insensitive to the accumulation of data in the sense that its decision bound remains constant for all sizes of a binomial sample. As a consequence, for a certain range of observed proportions, the PIC cannot favor one of the two models even when the data reasonably support it, no matter how large a sample is collected. In a subset of the cases, it can be shown that the unconstrained model is never selected with any supportive data of any size since ΔPIC is signed towards the constrained model over all possible sample proportions. By contrast, in all of these testing scenarios, the Bayes factors perform as expected, selecting the model that receives sensible support from the data. The selection based on Bayes factors is also consistent, recovering the underlying process as the sample size increases.

In this section, we examine distinctive properties of the PIC in more general cases. It turns out that the same behavior that hallmarks the PIC when applied to binomial models holds for the general exponential family distributions, which include many standard probability distributions in statistics, such as Bernoulli, binomial, Poisson, normal, and so forth (Schervish, 1995). When the PIC is applied to Models 1 and 2 constructed by imposing two different priors on an exponential family, $\Delta\text{PIC} = \text{PIC}_1 - \text{PIC}_2$ has the form

$$\Delta\text{PIC} = 2\left(\sum_{i=1}^{n} T(y_i)\right)\left[\int \eta(\theta_2)p_2(\theta_2)\,\mathrm{d}\theta_2 - \int \eta(\theta_1)p_1(\theta_1)\,\mathrm{d}\theta_1\right] - 2n\left[\int A(\theta_2)p_2(\theta_2)\,\mathrm{d}\theta_2 - \int A(\theta_1)p_1(\theta_1)\,\mathrm{d}\theta_1\right], \tag{6}$$

where $T(y)$ is a sufficient statistic for the model parameter θ, $\eta(\theta)$ is a function of θ (called the natural parameter), $A(\theta)$ is the logarithm of a normalizing constant, and $p_1(\theta_1)$ and $p_2(\theta_2)$ are prior distributions representing different hypotheses about θ (possibly on subdimensions of θ, hence subscripts 1 and 2 on θ). The observations, $y_1, \ldots, y_n$, are independently and identically distributed, under the condition of which $\sum_{i=1}^{n} T(y_i)$ becomes a sufficient statistic made of all observations.[2]

The above form of ΔPIC, of which all the expressions of ΔPIC in Table 1 for binomial model testing are special cases, shows

[2] This expression extends to the case of a vector parameter in a straightforward fashion. In such a case, the first term of ΔPIC in Eq. (6) is replaced by a linear combination of multivariate sufficient statistics.

that the criterion is a linear function of the model's sufficient statistics. It also manifests that the decision bound, or the sufficient statistic normalized by the sample size n (i.e., $\frac{\sum_{i=1}^{n} T(y_i)}{n}$) solving $\Delta\text{PIC} = 0$, remains constant for all n's. Therefore, in its applications to hypothesis testing, we expect precisely the same behavior as demonstrated with binomial models to arise: There exist a range of observed statistics that cannot evidence one of the models, no matter how strong evidence is accumulated in the statistic in support of the model.

We believe that the PIC's problematic selection extends to models outside exponential families, but its specific form is uncertain. Nonetheless, some insight can be gained by considering the following relationship:

$$\begin{aligned}
\text{PIC} &= -2\int \log[f(y|\theta)]p(\theta)\mathrm{d}\theta \\
&= -2\int \log\left[\int f(y|\theta)p(\theta)\mathrm{d}\theta \cdot \frac{f(y|\theta)p(\theta)}{\int f(y|\theta)p(\theta)\mathrm{d}\theta} \cdot \frac{1}{p(\theta)}\right] p(\theta)\mathrm{d}\theta \\
&= -2\int \log\left[\int f(y|\theta)p(\theta)\mathrm{d}\theta \cdot \frac{p(\theta|y)}{p(\theta)}\right] p(\theta)\mathrm{d}\theta \\
&= -2\log\int f(y|\theta)p(\theta)\mathrm{d}\theta + 2\int\left[\log\frac{p(\theta)}{p(\theta|y)}\right]p(\theta)\mathrm{d}\theta \\
&= -2\log \text{ML} + 2D_{\text{KL}}\left(p(\theta) \parallel p(\theta|y)\right),
\end{aligned} \tag{7}$$

where $D_{\text{KL}}(p(\theta) \parallel p(\theta|y))$ is the Kullback–Leibler (KL) divergence of the posterior distribution from the prior distribution. This means that the PIC can be regarded as a two-part decomposition: (the negative logarithm of) the model's marginal likelihood and the KL divergence between prior and posterior distributions. With two models under comparison, ΔPIC becomes

$$\begin{aligned}
\Delta\text{PIC} = &-2\log\text{BF} + 2\left[D_{\text{KL}}\left(p_1(\theta) \parallel p_1(\theta|y)\right)\right. \\
&\left. - D_{\text{KL}}\left(p_2(\theta) \parallel p_2(\theta|y)\right)\right].
\end{aligned} \tag{8}$$

This result translates that the selection outcome of ΔPIC deviates from that of $-2\log$BF depending on the difference of two KL divergence terms, which indicate the relative degree of departure of each model's posterior distribution from their priors. Typically, the posterior of a model changes from a widely dispersed prior to an increasingly peaked, limiting distribution as the sample size increases. The KL divergence in the above expressions as a function of n can be considered to measure the *rate* of such convergence. In fact, the convergence rate of Bayesian posteriors under various conditions, not just the fact of convergence, is an ongoing research topic and the KL divergence of a posterior plays a key role in such analysis (Ghosal, Ghosh, & van der Vaart, 2000; Kleijn & van der Vaart, 2006). Results relevant to the present discussion concern two distinct conditions: The model is well-specified versus mis-specified (i.e., the model contains the underlying, data-generating process vs. it does not). It has been shown that the posterior of a well-specified model converges to its limiting distribution at an optimal rate under mild regularity conditions, whereas the same rate for a mis-specified model is achieved when more stringent conditions are met. Recall that the PIC exhibits clearly counterintuitive selection behavior when the unconstrained model is well-specified yet the constrained model is mis-specified (i.e., the truth lies only in the unconstrained model). This behavior may be put in perspective when the deviation of the PIC from the BF shown above is considered together with the aforementioned findings about posterior convergence rates: the PIC can penalize the well-specified, unconstrained model more severely than the BF does because the convergence of the unconstrained model's posterior is faster than the constrained model's, magnifying the KL divergence term in the PIC as a penalty.

General results about exact conditions in which the PIC produces selection outcomes that are distinct from the BF's for the above reason are not available. Instead, we may consider a special case in which the PIC's deviation from the BF takes a simplest form: when a point hypothesis, in which all parameters are fixed in a model, is tested against a larger model with (some of) the parameters set free. An example is the test of an equality constrained binomial model (M_c) against the unconstrained model (M_{01}) illustrated earlier (results shown in the bottom row of Fig. 2). In this case, the KL divergence in Eq. (8) is always zero for the constrained model (Model 2) because its prior and posterior are an identical point mass, whereas the KL divergence for the unconstrained model (Model 1) becomes positive as its posterior differs from the prior with data accumulation. This makes the difference of the two KL divergences in Eq. (8) strictly positive, leading to $\Delta\text{PIC} > -2\log\text{BF}$. Consequently, when compared to the point hypothesis, the unconstrained model tends to be favored less under the PIC than under the BF. This observation certainly extends to general cases. That is, with the PIC, an unconstrained, well-specified model can be penalized over the mis-specified, point hypothesis relative to the case with the BF.

3. Discussion

The present paper concerns a newly proposed method of statistical model selection, the prior information criterion (van de Schoot et al., 2012), and reports an analysis of its behavior when applied to test an inequality- or an equality-constrained hypothesis. The results show that the PIC yields puzzling outcomes of model selection, which are demonstrated in examples of testing binomial models under various constraints, and explained using analytic derivations of the PIC in more general settings. In sum, it was found that there exists a situation in which one of the models under comparison cannot be selected by the PIC even when the data sensibly support that model, no matter how much data are collected. Specifically, it is possible that the model chosen by the PIC does not contain the underlying truth whereas the alternative model does.[3] In the same situation, the Bayes factor favors the model that receives sensible support from the data. As predicted in general cases (Schwartz, 1965), the selection behavior of the Bayes factor is consistent, favoring the correct, data-generating model as data accumulate.

Inconsistent selection behavior is in fact a large-sample property of some existing model selection criteria, most often associated with the Akaike information criterion (AIC; Akaike, 1973). One may wonder if the inconsistency of the PIC can be understood in a similar way. There are critical differences, however, in what constitutes "inconsistency" between the AIC and the PIC. The AIC is inconsistent in the sense that it is possible for a constrained model not to be recovered with a large sample when the true distribution lies both in the constrained and the alternative, encompassing model (Bozdogan, 1987). This selection is viewed as being inconsistent because, in such a case, the smaller, nested model should be considered closer to the truth than the larger model (i.e., inconsistent with respect to the order of model classes in their overall discrepancy from the true distribution). This outcome is far from being puzzling and is defended by the fact that the AIC is designed to estimate the fitted model's KL divergence from the true distribution, not the order of compared model classes in the aforementioned sense. By contrast, the PIC does not let one select a model even when the truth lies *only* in that model and not in

[3] In fact, an example of how the PIC can lead to a problematic inference of this kind is already present (but neither observed nor discussed) in van de Schoot et al. (2012). For example, when data were simulated from a population where $\mu_1 = -1$ and $\mu_2 = 1$, the PIC preferred the equality constrained model ($\mu_1 = \mu_2$) over the unconstrained model (μ_1, μ_2), despite the fact that the unconstrained, but not the constrained model, includes the truth (see their Figure 3 and Mulder, 2014).

the alternative one. Furthermore, the PIC in certain cases does not recover the true model with data of all sample sizes. To the best of our knowledge, there are no existing model selection criteria that exhibit such behavior.

Methods for testing hypotheses statistically, or selecting among competing statistical models, are important tools for scientific research. Development of a new method should be welcomed as there must not be a single, absolute criterion for such testing. The PIC would have been a valuable addition to the existing methods. However, based on our analyses reported in the current paper, we must express reservations about its use.

Acknowledgments

We would like to thank several anonymous reviewers and Michael Lee for their helpful comments on previous versions of this paper.

Appendix A. Supplementary materials

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jmp.2017.06.002.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *B. Petrov, & B. Csaki (Eds.), Second international symposium on information theory* (pp. 267–281). Academiai Kiado, Budapest.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370. http://dx.doi.org/10.1007/BF02294361.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511790485.

Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilités et ses applications*, *13*, 23–27.

Ghosal, S., Ghosh, J. K., & van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, *28*, 500–531. http://dx.doi.org/10.1214/aos/1016218228.

Jeffreys, H. (1961). *Theory of probability*. Oxford: Oxford University Press. http://dx.doi.org/10.1093/gji.6.4.555.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795. http://dx.doi.org/10.1080/01621459.1995.10476572.

Kleijn, B. J. K., & van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, *34*, 837–887.

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics & Data Analysis*, *71*, 448–463. http://dx.doi.org/10.1016/j.csda.2013.07.017.

Schervish, M. J. (1995). *Theory of statistics*. New York: Springer.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift fur Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *4*, 10–26.

van de Schoot, R., Hoijtink, H., Romeijn, J.-W., & Brugman, D. (2012). A prior predictive loss function for the evaluation of inequality constrained hypotheses. *Journal of Mathematical Psychology*, *56*, 13–23. http://dx.doi.org/10.1016/j.jmp.2011.10.001.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498. http://dx.doi.org/10.1016/j.jmp.2010.07.003.