An Objective Approach to Prior Mass Functions for Discrete Parameter Spaces

C. Villa and S. G. Walker

We present a novel approach to constructing objective prior distributions for discrete parameter spaces. These types of parameter spaces are particularly problematic, as it appears that common objective procedures to design prior distributions are problem specific. We propose an objective criterion, based on loss functions, instead of trying to define objective probabilities directly. We systematically apply this criterion to a series of discrete scenarios, previously considered in the literature, and compare the priors. The proposed approach applies to any discrete parameter space, making it appealing as it does not involve different concepts according to the model. Supplementary materials for this article are available online.

KEY WORDS: Binomial; Hypergeometric; Kullback–Leibler divergence; Loss function; Objective prior.

1. INTRODUCTION AND BACKGROUND

The purpose of this article is to outline a novel approach to constructing objective prior distributions for discrete parameter spaces. The difficulty in dealing with this type of parameter space is well known and, while problem-specific solutions have been successfully developed, it seems that a generalized approach that could deal with any type of problem remains elusive. In recent work, Berger, Bernardo, and Sun (2012) presented a solution to the problem by embedding the discrete structure in a continuous one, through the application of one or more of four possible approaches, and applying reference analysis to the continuous structure to obtain a prior distribution. In this cited article, it is also possible to find literature to past work on the development of a general procedure to obtain objective priors for discrete parameter spaces. Other work on the subject can be found, for example, in Rissanen (1983), Lindsay and Roeder (1987), and Barger and Bunge (2008).

Our aim is to propose a general framework that can be applied to any discrete parameter space, and which is objective in the sense that, once a model has been chosen, it does not require subjective input. Alongside this, we intend to define prior distributions for all the models covered in Berger, Bernardo, and Sun (2012), from now on referred to as BBS. It is worth mentioning that we have kept the notation as much as possible similar to the one in BBS.

1.1 Alternative Approaches

We would first like to briefly summarize the approach outlined by BBS. It is in fact our intention to derive prior distributions for the discrete models discussed in their article, and any comparison we make largely benefits from this short review.

Their basic idea is to embed the original model, defined by a discrete parameter (or vector of discrete parameters), into a continuous model, such that the structure is preserved. Then, prior distributions are obtained by applying the regular reference methodology to the continuous setup. Literature on reference priors can be found, for example, in Bernardo (1979, 2005), Bernardo and Smith (1994), Berger and Bernardo (1992), and Berger, Bernardo, and Sun (2009).

Embedding a discrete model into a continuous one can be performed in several ways. The authors have identified four ways, as it appears that there is no unique approach that can be applied to any discrete scenario; in particular, among the ones examined in the article. These four approaches are as follows.

**Approach 1.** Here, the discrete parameter is treated as a continuous parameter. The reference prior can therefore be directly found on this parameter. This approach can be applied to the problem of estimating the population size for a Type II censoring experiment, or to a hypergeometric model, for example.

**Approach 2.** This relies on the fact that, sometimes, it is possible to add a hierarchical level of modeling to create a continuous hyperparameter. This is feasible and natural, and a reference prior can be defined on the hyperparameter. However, the hierarchical level is not necessarily unique, different objective priors can be defined for the same problem. This approach can be applied to the hypergeometric and multivariate hypergeometric models.

**Approach 3.** Here, the asymptotic distribution of a (consistent) estimator for the parameter of interest is considered. A reference prior is then derived for the parameter in the asymptotic distribution, pretending it is continuous. This property was discussed by Berger and Bernardo (1992) and is applied to the hypergeometric, binomial-beta, and binomial models.

**Approach 4.** Here, a limiting operation on the parameter is derived, so that the problem can be seen as continuous by letting the parameter itself go to infinity. Thus, a reference prior is applied. This approach can be used to derive prior distributions for the hypergeometric, binomial-beta, and binomial models.

A more detailed exposition of these four approaches, and reference to appropriate literature, can be found in BBS.
1.2 The Idea

Let us consider a probability distribution \( f(x \mid \theta) \), which depends only on the value of a discrete parameter \( \theta \in \Theta \). In a subjective Bayesian approach, the prior \( \pi(\theta) \) represents the initial degree of belief that we have about the possible values that \( \theta \) can take within the parameter space. Then, by combining it with the information contained in the observed data, represented by the likelihood function, the initial beliefs are updated to the posterior probability distribution \( \pi(\theta \mid x) \). The prior and the posterior retain the same meaning.

While it is problematic to assign an objective probability, while attempting to be least informative, we argue that the fact that parameter value \( \theta \) has been selected to be a part of the model conveys information that can be exploited. For such inclusion, the value \( \theta \) must have a worth associated with it. We can objectively evaluate the worth by determining what is objectively lost if the \( \theta \) value is removed from the model.

Let us denote by \( \pi(\theta) \) the prior distribution for the discrete parameter \( \theta \in \Theta \). If a prior \( \pi \) has been assigned, then we can link this to a worth of each element by means of the self-information loss function \( - \log \pi(\theta) \); or, equivalently, the utility of \( \theta \) is denoted by \( \log \pi(\theta) \). Details of this particular loss/utility function can be found, for example, in Merhav and Feder (1998). We can then identify an appropriate objective way to associate a utility with each \( \theta \), representing its worth in the model, and the prior distribution \( \pi(\theta) \) then follows. Furthermore, we note that in this way the Bayesian approach is conceptually consistent, as we update an initial (i.e., prior) utility assigned to \( \theta \) based on the observed data, represented by the likelihood function, the initial beliefs are updated to the posterior utility distribution \( \pi(\theta \mid x) \). The prior and the posterior retain the same meaning.

We want \( u_1(\theta) \) and \( u_2(\theta) \) to be matching utility functions; though as it stands \( -\infty < u_1 \leq u_2 < \infty \), and we want \( u_1(\theta) = -\infty \) when \( u_2(\theta) = 0 \). The scales are matched by taking exponential transformations; so \( \exp(u_1) \) and \( \exp(u_2) - 1 \) are on the same scale. Hence, we have

\[
e^{u_1(\theta)} = \pi(\theta) \propto e^{u_2(\theta)},
\]

where

\[
g(u) = \log(e^u - 1).
\]

By setting the functional form of \( g \) in (1) as it is defined in (2), we derive the objective prior distribution for the discrete parameter \( \theta 

\[
\pi(\theta) \propto \exp \left\{ \min_{\theta' \neq \theta \in \Theta} (D_{KL}(f(\cdot \mid \theta) \mid f(\cdot \mid \theta'))) \right\} - 1.
\]

In other words, for each value \( \theta \in \Theta \) for which the divergence measured between the two defined models \( f(x \mid \theta) \) and \( f(x \mid \theta') \) is minimum; the prior mass at \( \theta \) is proportional to the obtained Kullback–Leibler divergence. The prior defined in (3) will be applied to all the models considered in this article.

The outline of the article follows the sections in BBS. So in Sections 2 to 6, we define the objective prior in accordance with the approach we propose and compare the results with the ones obtained by BBS. Each of this sections has the same structure: first, we examine the model and the prior distribution defined by BBS; we then derive the prior applying (3) and compare the two. Section 7 contains final discussion points.

2. A POPULATION SIZE MODEL

The first case considered is the estimation of the size of a population by means of a Type II censoring. In this experiment, there is a sample of \( N \) units with a lifetime that is modeled by an exponential distribution with unknown parameter \( \lambda \). The experiment stops when a predetermined number of failures \( R \) is reached. Failure times \( t_1 \leq \cdots \leq t_R \), which are assumed to be independent, have the following joint distribution:

\[
p(t_1, \ldots, t_R \mid N, \lambda) = \frac{N!}{(N - R)!} \lambda^R \exp \left\{ -\lambda [t_1 + \cdots + (N - R)tr] \right\}, \quad N \geq R.
\]

It can be shown (Goudie and Goldie 1981) that inference on \( N \) is equivalently performed by considering the following density:

\[
p(V \mid N) = \frac{1}{(R - 2)! (N - R)!} \frac{1}{(N + V - R)!} g_R(V), \quad 1 < V < R,
\]

where \( V = (t_1 + \cdots + t_R)/t_R \) and

\[
g_R(V) = \sum_{i=1}^{[V]} (-1)^{i-1} \binom{R - 1}{i - 1} (V - i)^{R - 2}.
\]

BBS applied Approach 1. That is, the parameter space of \( N, \{R, R + 1, \ldots\} \) is embedded in the interval \( (R - 0.5, \infty) \), and the reference prior for \( N \) is obtained as \( \pi(N \mid R) \propto \sqrt{t_R(V)} \),
where \( i_R(N) \) is the Fisher information. The prior for \( \theta = N - R + 1 \), given \( R = 2, 3, 4 \), is 

\[
\pi(\theta \mid R) = \begin{cases} 
\frac{1}{(\theta+1)!} & \text{if } R = 2, \\
1.3036 \frac{\sqrt{\theta(\theta+2\theta+1)}}{(\theta+1)(\theta+2)} & \text{if } R = 3, \\
1.6017 \frac{\sqrt{\theta(\theta+3\theta+2\theta+1)}}{(\theta+1)(\theta+2)(\theta+3)} & \text{if } R = 4.
\end{cases}
\]

Numerically, BBS verified that the prior is proper up to \( R = 100 \), which is necessary as the likelihood function converges to a constant when \( N \) tends to infinity, and the posterior would be improper. Furthermore, the prior distributions for different values of \( R \) are all very similar, except at \( \theta = 1 \). Therefore, the recommendation is to use the prior for \( R = 2 \) for any value of \( R \), as it is considered a good approximation.

### 2.1 The Objective Prior

To define the prior distribution for \( N \), by applying the objective approach we propose, we first consider the Kullback–Leibler divergence between two densities of type (4) with different value of \( N \):

\[
D_{KL}(p(V \mid N)\|p(V \mid N + c)) = \int p(V \mid N) \log \left( \frac{p(V \mid N)}{p(V \mid N + c)} \right) dV
\]

\[
= \log \left( \frac{N!}{(N + c - R)!} \right) + R \mathbb{E} \left[ \log \left( \frac{V + N + c - R}{V + N - R} \right) \right],
\]

where the expectation is with respect to \( p(V \mid N) \), and \( c \) is an integer. This is an increasing function in \( c \), which means that the nearest model to \( p(V \mid N) \), in terms of Kullback–Leibler divergence, is either \( p(V \mid N - 1) \) or \( p(V \mid N + 1) \). We have ascertained by computation that the Kullback–Leibler divergence from \( p(V \mid N) \) is minimized by \( c = 1 \), rather than \( c = -1 \). Therefore, by applying (3), we see that the objective prior for \( N \) is given by

\[
\pi(N \mid R) \propto \exp \{ D_{KL}(p(V \mid N)\|p(V \mid N + 1)) \} - 1 \propto \frac{N + 1 - R}{N + 1} \exp \left[ R \mathbb{E} \left[ \log \left( \frac{V + N + 1 - R}{V + N - R} \right) \right] \right] - 1. \tag{5}
\]

The prior distribution (5) is proper, as the following theorem, which has proof in the online supplementary material, shows.

**Theorem 1.** Consider the density \( p(V \mid N, R) \) defined in (4), with \( N \geq R = \{1, 2, \ldots \} \) and \( V \in (1, R) \). Then the prior distribution \( \pi(N \mid R) \) is proper.

### 2.2 Illustrations and Comments

The prior distribution for \( N \) assigns large mass for small values of the parameter and rapidly decreases as \( n \) increases. This is graphically verifiable in Figure 1 (right graph), where the priors for \( R = 2, 3, 4 \) have been plotted. This behavior is similar to the one obtained by BBS (left graph). It has to be noted that we have analytically proved that our prior is proper. Given that the likelihood function is constant as \( N \to \infty \), as discussed in BBS, this is a necessary condition for the posterior to be proper. We have performed an analysis of the posterior performance (not shown here) by considering the approximate frequentist coverage of the 95% credible intervals. In addition, we have considered the relative mean squared error, \( \sqrt{\text{MSE}(\theta)}/\theta \), for the posterior means. For 20,000 simulations of sample size \( m = 1 \), the coverage of BBS prior and our prior are remarkably similar although conservative, as they tend to be higher than the nominal level. In terms of MSE, BBS prior has a better performance for low values of \( \theta \) (except for \( \theta = 1 \)); however, when \( \theta \) increases, the two priors tend to have similar performance. Our prior has a more stable performance, in terms of MSE, over the whole parameter space considered. We have performed the simulations for \( m = 1 \) so the differences in the frequentist performances of the priors are more prominent. However, for more realistic sample sizes (e.g., \( m = 30 \)), the performances become remarkably similar.

### 3. HYPERGEOMETRIC MODEL

Let us consider now a hypergeometric distribution with probability mass function given by

\[
p(r \mid N, R, n) = \binom{R}{r} \binom{N - R}{n - r}, \quad r \in \{\max(0, n - (N - R)), \min(n, R)\},
\]

with the population size \( N \) and the sample size \( n \) known, and \( R = 0, 1, \ldots, N \), representing the units in the population, which satisfy a certain criterion (or that possess a certain characteristic). The parameter \( R \) is unknown, and the aim is to define the prior \( \pi(R) \).

BBS obtained a prior on \( R \) by applying Approach 2. It is assumed that the parameter has a binomial hierarchical model, \( \text{Bin}(R \mid N, p) \), where the continuous parameter \( p \) is unknown. Therefore, the problem reduces in finding the reference prior for \( p \). As shown in Bernardo and Smith (1994), this is accomplished by marginalizing out the parameter in the lower level. Given that the reference prior for the parameter \( p \) of a binomial distribution is the Jeffreys’ rule prior, we have

\[
\pi(R \mid N) = \int_0^1 \text{Bin}(R \mid N, p) \text{Be}(p \mid 1/2, 1/2) dp = \frac{1}{\pi} \frac{\Gamma(R + 1/2)\Gamma(N - R + 1/2)}{\Gamma(R + 1)\Gamma(N - R + 1)}, \quad R = 0, 1, \ldots, N,
\]

which, as it is defined on a finite parameter space, is proper.

### 3.1 The Objective Prior

Let us consider two hypergeometric models that differ for the value of parameter \( R \) only, say \( p_\text{L} = p(r \mid N, R, n) \) and \( p_\text{R} = p(r \mid N, R', n) \). The Kullback–Leibler divergence between the two models is given by

\[
D_{KL}(p_\text{L} \| p_\text{R}) = \sum_{r \in R} \left[ \frac{\binom{R}{r} \binom{N - R}{n - r}}{\binom{N}{n}} \log \left( \frac{\binom{R}{r} \binom{N - R}{n - r}}{\binom{R'}{r} \binom{N - R'}{n - r}} \right) \right] \left( \frac{\binom{N}{n}}{\binom{N - R}{n - r}} \right) \left( \frac{\binom{N - R}{n - r}}{\binom{N - R'}{n - r}} \right)
\]

\[
= \log \left( \frac{R!}{R'} \left( \frac{N - R'}{N - R} \right) \right).
\]
by swapping the role of the units that satisfy the criterion, we computed for $N$ to be known. If we indicate by $V$ the set of possible values of $n$, we have $V = \{0, n - (N - R)\}$, $n \in \{0, n - (N - R)\}$. To prove the above result, it is sufficient to rearrange the terms of Equation (6). In other words, the model with parameter $R$ is equal to the model with parameter $N - R$, for the same values of $N$ and $n$.

For a hypergeometric set of models with common known parameters $N$ and $n$, the minimum divergence from each element of the set is determined by the following Lemma 1, which has a proof in the online supplementary material.

**Lemma 1.** Consider the hypergeometric distribution $p_{R_0}$, with parameters $R_0$, $N$, and $n$, where $N$ and $n$ are assumed to be known. If we indicate by $p_R$ the hypergeometric distribution that differs from $p_{R_0}$ only by the number of units in the population $N$, which satisfy a certain criterion (i.e., $R_0$), then the Kullback–Leibler divergence from $p_{R_0}$ to $p_{R}$ is minimum when $R = R_0 + 1$, if $R_0 < N/2$, and $R = R_0 - 1$ if $R_0 > N/2$. If $R_0 = N/2$, then $D_{KL}(p_{R_0} || p_{R_0+1}) = D_{KL}(p_{R_0} || p_{R_0-1})$.

In deriving the objective prior for the parameter $R$ of the hypergeometric distribution, we make first the following considerations. We assume that parameters $N$ (population size) and $n$ (sample size) are known. Given the result of Lemma 1, for $R \leq N/2$ the minimum Kullback–Leibler divergence is obtained from (7) by setting $R' = R + 1$:

$$D_{KL}(p_R || p_{R+1}) = \sum_{r \in \mathcal{R}} \left[ \frac{\binom{R}{r} \binom{N-R}{n-r}}{\binom{N}{n}} \log \left( \frac{\binom{R}{r} \binom{N-R}{n-r}}{\binom{R+1}{r} \binom{N-(R+1)}{n-r}} \right) \right] = \log \left( \frac{N - R}{R + 1} \right) + \mathbb{E} \left[ \log \left( \frac{R + 1 - r}{N - R - n + r} \right) \right].$$

Therefore, for $R \leq N/2$, the prior is obtained by applying (3), and is given by

$$\pi(R \mid N, n) \propto \left( \frac{N - R}{R + 1} \right)^{n} \exp \left[ \mathbb{E} \left[ \log \left( \frac{R + 1 - r}{N - R - n + r} \right) \right] \right] - 1.$$

By symmetry, the prior mass for $R$, when $R > N/2$, is given by

$$\pi(R \mid N, n) = \pi(N - R \mid N, n).$$

### 3.2 Illustrations and Comments

To illustrate the behavior of the prior distribution as defined in (8) and (9), we can inspect Figure 2. The continuous curve shows the prior $\pi(R \mid N)$ for $N = 25$. The distribution has higher value at the extremes of the parameter space, symmetrically decreasing when approaching the center of the space (i.e., $N/2$). In the same graph, we see the prior obtained by BBS (dashed curve). Our prior is flatter than the one BBS proposed, but it has a very similar shape. In particular, it is symmetric about $N/2$.

Both we and BBS have computed the $\pi(R \mid N)$ for different values of $N$. In all circumstances, the distributions are similar, in terms of behavior. Unlike the objective prior of BBS, our prior depends on the value of $n$. 

![Figure 1](image1.png)

*Figure 1. Prior distribution for the population size model with BBS approach (left graph) and our approach (right graph). The prior has been computed for $R = 2, 3, 4$, and put on $\theta = N - R + 1$.**
As in Section 2, we have analyzed the frequentist performance of the posterior for 20,000 simulations (not showed here). In particular, for values of $N = 10$, $N = 20$, and $N = 25$, we have computed the (approximate) frequentist coverage of the 95% credible intervals for $n = 1$ and $n = 3$. Both BBS and our prior have similar coverage (although conservative), for $n = 1$ and $n = 3$. We have computed the MSE from the mean in all the above cases; the performance of BBS prior appears to be better than ours toward the center of the parameter space, while the roles are swapped in the extreme regions of the space. However, these differences are not substantial, and sensibly diminish for either $n > 1$ or $N > 10$.

4. MULTIVARIATE HYPERGEOMETRIC MODEL

Consider the multivariate hypergeometric distribution $MH_d(N, \mathbf{R}, n)$ of dimension $d$, with probability mass function:

$$p(\mathbf{r} \mid N, \mathbf{R}, n) = \frac{\binom{R_j}{r_j} \cdots \binom{R_d}{r_d}}{\binom{N}{n}}, \quad \mathbf{r} \in \mathbb{N}^d,$$

(10)

where $\mathbb{N}^d$ is the $d$-dimensional space of nonnegative integers, and with $n \in \{0, 1, \ldots, N\}$, $\sum_{j=1}^d R_j = N$, $\sum_{j=1}^d r_j = n$, and $r_j \leq \min(n, R_j)$ for $j = 1, \ldots, d$. For $d = 2$, we obtain the univariate hypergeometric distribution, discussed in Section 3. We assume that parameters $N$ and $n$ are known, and $\mathbf{R} = (R_1, \ldots, R_d)$ represents the vector of unknown parameters.

The prior distribution for vector $\mathbf{R}$ defined by BBS is based on Approach 2, similarly to Section 3, for the univariate hypergeometric model. In particular, BBS considered the vector of hyperparameters $\mathbf{p}_d$, where each element is defined as $p_i = R_i/N$, for $i = 1, \ldots, d$. This vector represents the hyperparameter vector of the multinomial distribution $MH_d(\mathbf{p}_d \mid N, \mathbf{p}_d)$, on which the prior mass is put on. This prior has the following form:

$$\pi_d(\mathbf{p}_d) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)} \prod_{j=1}^{d+1} p_j^{1/2 - 1},$$

(11)

which corresponds to the Jeffreys’ prior (Jeffreys 1961).

4.1 The Objective Prior

The Kullback–Leibler divergence between the multivariate hypergeometric distribution with parameters $N$, $\mathbf{R}$, and $n$, indicated by $p_{N, \mathbf{R},n}$, and the multivariate hypergeometric distributions with parameters $N$, $\mathbf{R} + \mathbf{a}$, and $n$, indicated by $p_{N, \mathbf{R} + \mathbf{a},n}$, where $\mathbf{a} \in \mathbb{Z}^d$, is given by

$$D_{KL}(p_{N, \mathbf{R},n} \parallel p_{N, \mathbf{R} + \mathbf{a},n})$$

$$= \sum_{\mathbf{r}} \left[ \log \frac{\left( \prod_{j=1}^d \binom{R_j}{r_j} \right) / \binom{N}{n}}{\prod_{j=1}^d \binom{R_j + a_j}{r_j} / \binom{N}{n}} \right]$$

$$= \sum_{j=1}^d \log \left( \frac{R_j!}{(R_j + a_j)!} \right) + \mathbb{E} \left[ \log \left( \frac{(R_j - r_j + a_j)!}{(R_j - r_j)!} \right) \right],$$

(12)

where $\mathbb{E}$ is the expectation of $\log \left( p_{N, \mathbf{R},n} / p_{N, \mathbf{R} + \mathbf{a},n} \right)$ with respect to $p_{N, \mathbf{R},n}$. The following lemma, which has a proof in the online supplementary material, determines the minimum Kullback–Leibler divergence from model $p_{N, \mathbf{R},n}$.

Lemma 2. Consider the $d$-dimensional multivariate hypergeometric distribution $p_{N, \mathbf{R},n}$, where parameters $N$ and $n$ are assumed to be known, with probability mass function as specified in (10). If we consider the hypergeometric distribution $p_{N, \mathbf{R}',n}$, which differs from $p_{N, \mathbf{R},n}$ by the composition of the unknown $d$-dimensional parameter vector $\mathbf{R}'$, then the Kullback–Leibler divergence between $p_{N, \mathbf{R},n}$ and $p_{N, \mathbf{R}',n}$ is minimum, when $\mathbf{R}' = \mathbf{R} + \mathbf{c}$, where $\mathbf{c}$ is a vector of dimension $d$ with $d - 1$ zeroes and, in correspondence of the element of $\mathbf{R}$ closer to $N/2$, has a minus or plus one depending if the “closeness” is, respectively, from above or below $N/2$.

To summarize the results obtained in the proof of Lemma 2, and to generalize to any multivariate hypergeometric distribution, we have that the smallest difference between $p_{N, \mathbf{R},n}$ and $p_{N, \mathbf{R}',n}$ is obtained when only one of the components of $\mathbf{R}$ is changed. In particular, when the change is an increase or decrease of one unit. Therefore, $\mathbf{c}$ will have $d - 1$ elements equal to zero and the remaining one is equal to plus or minus one. From the analysis of the Kullback–Leibler divergence between two bivariate hypergeometric models, the nearest model to $p_{N, \mathbf{R},n}$.
corresponds to the model \( p_{N,R+c.n} \), where \( c \) will have \( d-1 \) null elements and the remaining one, in position \( i \) (where \( i \) is the index of element \( R_i \) of \( R \) nearest to \( N/2 \), either from below or above), and will have value one if \( R_i \leq N/2 \), and minus one if \( R_i \geq N/2 \).

The objective prior for the \( R \) can be found from the result of the previous paragraph, and it has the form

\[
\pi(R) \propto \exp \left\{ D_{KL}(p_{N,R+n} \| p_{N,R+a,n}) \right\} - 1,
\]

where vector \( a \) is determined so that the divergence is minimized, as discussed in Lemma 2.

4.2 Illustrations and Comments

We have computed the prior distribution in the case of a bivariate hypergeometric distribution. In particular, we have considered a bivariate distribution with population size \( N = 10 \) and sample size \( n = 3 \). Table 1 represents the normalized prior distribution \( \pi(R) \), where \( R = (R_1, R_2) \). To have a better feeling, we have also plotted the distribution in Figure 3.

This prior can be compared with the one obtained by BBS, as in (11). We have computed the bivariate hypergeometric distribution considered in the illustration above, that is, for \( N = 10 \). This objective prior is represented in Table 2. Note that, due to the “new” nature of the problem, some areas of the discrete parameter space have now assigned a prior mass that tends to infinity.

If we compare our prior distribution with the one defined by BBS, we note some similarities. In particular, both show a symmetrical behavior, which can be noticed by row and by column, that is, by fixing one of the two parameters, the remaining is symmetrical around its central value.

5. BINOMIAL-BETA MODEL

Let us now consider a binomial model with parameters \( n \) and \( p \), with \( n = 1, 2, \ldots, d \) and \( p \in (0,1) \). We also assume that parameter \( p \) can be represented by a beta distribution with parameters \( a \) and \( b \), both strictly positive. The binomial-beta distribution is given by

\[
p(x | n) = \int_0^1 \frac{n!}{(n-x)!x!} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{x+a-1}(1-p)^{n-x+b-1} dp
\]

BBS derived the prior for \( n \) in the binomial-beta model by applying Approach 3 and Approach 4. As their choice, based on the analysis of frequentist properties of the posterior, is the one coming from Approach 3, we focus on that only. The approach aims to apply reference prior theory with a consistent estimator. In this case, the estimator for \( n \) is given by \( \hat{n} = (ak)^{-1}(a+b) \sum x_i \), where \( k \) is the number of independent samples from (13). Thus, the proposed prior is obtained by applying Jeffreys’ rule on the extended model \( p(\hat{n} | n) \approx N(\hat{n} | n, (n(n+a+b))/[a^2(a+b+1)k]) \), when \( n \) is considered continuous:

\[
\pi(n) \propto \frac{1}{\sqrt{n(n+a+b)}}.
\]

It is worth mentioning that BBS seemed to focus on values of \( a \) and \( b \) larger than one. In fact, to identify the best prior among the one deriving by applying Approach 3 and the one deriving by applying approach 4, they studied the frequentist coverage of credible intervals. And this comparison is done for values of \( a = b = 5, a = b = 20 \), and \( a = b = 50 \). As such, we consider the same restriction, that is, \( a, b > 1 \).

5.1 The Objective Prior

To find the model that is nearest, in terms of Kullback–Leibler divergence, to \( p(x | n) \), we first note that, as \( p(x = n | n') = 0 \) for \( n > n' \), we have

\[
D_{KL}(p(x | n) \| p(x | n')) = \infty \quad n > n'.
\]

Therefore, the nearest model will have the parameter representing the number of trials larger than \( n \). The Kullback–Leibler divergence from \( p_n \) to \( p_{n'} \) is minimum when \( n = n' + 1 \). This has been computationally verified. However, it can also be proved on the line of the proof of Lemma 3 in Section 6.

Thus,

\[
D_{KL}(p_n \| p_{n+1}) = \sum_{x=0}^{n} p_n \log \left( \frac{n!}{(n-x)!} \frac{\Gamma(n-x+b)(n+1-x)}{\Gamma(n+a-b)(n+1)!} \times \frac{\Gamma(n+1+a+b)}{\Gamma(n+1-x+b)} \right) = \log \left( \frac{n+a+b}{n+1} \right) + \mathbb{E} \left[ \log \left( \frac{n+1-x}{n-x+b} \right) \right],
\]

where the expectation is taken with respect to \( p_n \). The objective prior for \( n \) then is

\[
\pi(n) \propto \exp \left\{ D_{KL}(p_n \| p_{n+1}) \right\} - 1 = \frac{n+a+b}{n+1} \exp \left\{ \mathbb{E} \left[ \log \left( \frac{n+1-x}{n-x+b} \right) \right] \right\} - 1,
\]
and \( p \) from a binomial-beta distribution with parameters \( \pi \),

\[
\pi \propto \exp \left( \frac{1}{\sum_{x=0}^{n} \left( \frac{n + 1 - x}{n - x + b} \right)^{p_n}} \right) - 1.
\]

The following theorem, proved in the online supplementary material, shows that the posterior distribution for \( n \) is consistent.

**Theorem 3.** Consider the family of binomial-beta distributions \( p_n \), with \( n = 1, 2, \ldots \) and common parameters \( a \) and \( b \). We also assume that the true value of \( n \) is \( n_0 \). Given the prior distribution \( \pi(n) \) and the set of observations from \( p_{n_0}, x = x_1, \ldots, x_k \), the mass of the posterior corresponding to \( n_0 \) converges to one almost surely. That is,

\[
\pi(n_0) \mid n \geq n_0, x_1, \ldots, x_i \to 1,
\]

for \( k \to \infty \).

### 5.2 Illustrations and Comments

To have a feeling of the behavior of the prior in (15), we have computed it for different values of \( a \) and \( b \). In particular, as shown in Figure 4 (right graph), we have computed the prior for the parameter \( n \) given \( a = b = 5, a = b = 20 \), and \( a = b = 50 \). The three prior distributions are very similar. As expected, a large part of the mass is put on relatively small values of \( n \).

To compare our prior to the BBS one, we have computed the prior in (14) for \( n = 1, \ldots, 70 \) and \( a = b = 5, a = b = 20, a = b = 50 \). The results are in Figure 4 (left graph). Unlike...
Villa and Walker: An Objective Approach to Prior Mass Functions for Discrete Parameter Spaces

Figure 4. BBS objective prior (left graph) and our objective prior (right graph) for the Binomial-Beta model with \(a = b = 5\), \(a = b = 20\), and \(a = b = 50\).

to our approach, it appears that the values of \(a\) and \(b\) play a nonnegligible role, as the three distributions are different; in particular for small values of \(n\).

The comparison of our prior versus the one of BBS has been completed through a simulation study of the frequentist performance of the posterior. In particular, for each of the cases \(a = b = 5\), \(a = b = 20\), and \(a = b = 50\), we have performed 20,000 simulations of sample size \(m = 1\). As mentioned in Section 2.2, the choice of \(m = 1\) relies on the fact that possible differences are more prominent for small sample sizes. We have computed the frequentist coverage of the 95\% credible intervals and the frequentist relative MSE from the posterior means. Figure 5 shows the results of the simulation. On the left, we have the frequentist coverage, and on the right the relative MSE. Both priors lead to very similar coverage performance, with a more conservative value for \(n < 5\). In this area of the parameter space, our prior has a better MSE, with the difference from the BBS prior decreasing rapidly as \(n\) increases. For \(a = b = 5\), the results are consistent with the above; however, the differences in the MSE are more significant. For \(a = b = 50\), both priors have, substantially, the same performances.

It appears that the fact that our prior, compared to BBS’s, is less sensible to the values of \(a\) and \(b\) represents an advantage, as the simulation results show. This is particularly relevant for small values of the parameters.

6. BINOMIAL MODEL

Let us assume that the random variable \(x\) is binomially distributed with number of trials \(n\) and probability of success \(p\). Its probability mass function is

\[
f(x \mid n, p) = \binom{n}{x} p^x (1 - p)^{n-x} \quad x = 0, 1, \ldots, n.
\]

For simplicity in the notation, the model representing a binomial random variable \(x\), with parameters \(n\) and \(p\), will be written as \(f_x\) instead of \(f(x \mid n, p)\).

The prior for parameter \(n\) (with known \(p\)) defined by BBS has the following form:

\[
\pi(n \mid p) \propto \frac{1}{\sqrt{n}}. \tag{17}
\]

Figure 5. Frequentist coverage of 95\% credible intervals for \(n\) (left graph) and square root of the relative mean squared error for the mean (right graph). The simulation has been run for \(a = b = 20\) and for BBS prior (black-dashed line) and our prior (red-continuous line).
This prior has been obtained by applying Approach 3, as defined by the authors, and is based on the asymptotic properties (i.e., for large \( n \)) of the estimator:

\[
\hat{n} = \frac{\bar{x}_{k}}{p}, \tag{18}
\]

where at the numerator we have the mean of \( k \) independent replications from \( X \sim \text{Bin}(x \mid n, p) \). As already mentioned, this approach does not allow the use of efficient estimators, and in fact (18) is inefficient. It is worth mentioning that for this particular model, all four approaches considered by the authors were applicable. The choice has been made on the basis of the frequentist properties and the simplicity of (17).

### 6.1 The Objective Prior

Let us consider the following two binomial distributions: \( f_{n} \) and \( f_{n'} \), with \( n = 1, 2, \ldots, n' = 1, 2, \ldots \) and \( n \neq n' \), where \( n \) and \( n' \) represent the number of trials for each distribution. Furthermore, we assume that both distributions have the same value of the known parameter \( p \). The Kullback–Leibler divergence between \( f_{n} \) and \( f_{n'} \) is given by

\[
D_{KL}(f_{n} \parallel f_{n'}) = +\infty,
\]

if \( n' < n \); else:

\[
D_{KL}(f_{n} \parallel f_{n'}) = \sum_{x=0}^{n} \left[ f_{n} \log \left( \frac{f_{n}}{f_{n'}} \right) \right]
= \sum_{x=0}^{n} \left[ \left( \frac{n}{x} \right) p^{x}(1-p)^{n-x} \times \log \left( \frac{\left( \frac{n}{x} \right) p^{x}(1-p)^{n-x}}{\left( \frac{n'}{x} \right) p^{x}(1-p)^{n'-x}} \right) \right]
= E \left\{ \log \left( \frac{n}{x} \right) \right\} - E \left\{ \log \left( \frac{n'}{x} \right) \right\} + (n-n') \log(1-p),
\]

where \( E \) represents the expected value with respect to \( f_{n} \). The following lemma (which has proof in the online supplementary material) shows that the nearest model to \( f_{n_0} \) is \( f_{n_0+1} \).

**Lemma 3.** Consider the binomial distribution \( f_{n_0} \). If we indicate by \( f_{n} \), with \( n > n_0 \), the generic binomial distribution that differs from \( f_{n_0} \) only by the number of trials, then the Kullback–Leibler divergence from \( f_{n_0} \) to \( f_{n} \) is minimum when \( n = n_0 + 1 \). Thus, the minimum divergence from model \( f_{n} \) has the form:

\[
D_{KL}(f_{n} \parallel f_{n+1}) = \sum_{x=0}^{n} \left\{ \left( \frac{n}{x} \right) p^{x}(1-p)^{n-x} \log \left( \frac{n}{x} \right) \right\}
- \sum_{x=0}^{n} \left\{ \left( \frac{n}{x} \right) p^{x}(1-p)^{n-x} \log \left( \frac{n+1}{x} \right) \right\}
+ [n-(n+1)] \log(1-p)
= \sum_{x=0}^{n} \left\{ \left( \frac{n}{x} \right) p^{x}(1-p)^{n-x} \times \log \left( \frac{n!}{x!(n-x)!} \right) \right\} - \log(1-p)
\]

\[
= \sum_{x=0}^{n} \left\{ \log(n+1-x) \left( \frac{n}{x} \right) p^{x}(1-p)^{n-x} \right\}
- \log(n+1) - \log(1-p).
\]

Therefore, the prior distribution is given by

\[
\pi(n) \propto \frac{1}{(n+1)(1-p)} \exp \left\{ \sum_{x=0}^{n} \log(n+1-x) \left( \frac{n}{x} \right) \right\} - 1,
\]

that can alternatively be expressed in the form

\[
\pi(n) \propto \frac{1}{(n+1)(1-p)} \prod_{x=0}^{n} \left( n+1-x \right)^{(c)p^{x}(1-p)^{n-x}} - 1,
\]

which is useful to employ in some analytical procedures. The objective prior for \( n \) is improper. In fact, we have

\[
\sum_{n=1}^{\infty} \left\{ \frac{1}{(n+1)(1-p)} \prod_{x=0}^{n} \left( n+1-x \right)^{(c)p^{x}(1-p)^{n-x}} - 1 \right\}
\geq \sum_{n=1}^{\infty} \left[ \frac{1}{(n+1)(1-p)} - 1 \right], \tag{19}
\]

as \( \sum_{x=0}^{n} \left( n+1-x \right)^{(c)p^{x}(1-p)^{n-x}} \right) / (n+1-p) \geq 1 \). Therefore, as \( \sum_{n=1}^{\infty} \left[ 1/(n+1)(1-p) - 1 \right] = \sum_{n=1}^{\infty} [n/(n+1)(1-p)] = \infty \), the left-hand side of (19) diverges as well. Therefore, given that the prior distribution is improper, the following theorem, which has proof in the supplementary material, shows that the posterior is proper.

**Theorem 4.** Assume that we observe the data point \( x_{1} \) from a binomial distribution with parameters \( n \) and \( p \). Also, assume a prior distribution for the parameter \( n \) as \( \pi(n) \propto \exp \{ D_{KL}(f_{n} || f_{n+1}) \} \). Then, the posterior distribution is given by

\[
\pi(n \mid x_{1}) \propto \left[ \frac{1}{n+1}(1-p) \prod_{x=0}^{n} \left( n+1-x \right)^{(c)p^{x}(1-p)^{n-x}} - 1 \right]
\times \left( \frac{n}{x_{1}} \right)^{p x_{1}(1-p)^{n-x-1}}, \tag{20}
\]

is proper.

The consistency of the posterior distribution is examined in the following theorem, the proof of which is in the online supplementary material.

**Theorem 5.** Consider the family of binomial distributions \( f_{n} \), with \( n = 1, 2, \ldots \) and common parameter \( p \). Also assume that the true value of the parameter \( n \) is \( n_{0} \). Given the prior distribution \( \pi(n) \) and the set of observations from \( f_{n_{0}}, x = \{x_{1}, \ldots, x_{k} \} \), the mass of the posterior corresponding at \( n_{0} \) converges to 1 almost surely. That is,

\[
\pi(n_{0} \mid n \geq n_{0}, x_{1}, \ldots, x_{k}) \rightarrow 1, \tag{21}
\]
6.2 Illustrations and Comments

To have an understanding of the objective prior for $n$, we have computed and plotted it for a given value of $p = 0.5$, and for $n = 1, \ldots, 100$. The result is shown in Figure 6. We see that the highest mass is put on $n = 1$, as the largest divergence is $D_{KL}(f_1 \parallel f_2)$, and that it will decrease as $n$ increases. This is obvious, as the difference between a binomial with $n$ number of trials and the binomial with $n+1$ number of trials, diminishes as $n$ tends to infinity; as such, the Kullback–Leibler divergence measured between the two models tends to zero and so does the prior.

The comparison between the prior we propose and the prior defined by BBS is straightforward. By simply inspecting expression (17), we notice that the BBS prior distribution assigns large mass to small values of $n$, and then it decreases as $n$ becomes larger (Figure 6). In particular, the behavior is similar to our prior; even though our prior depends on the value of $p$, by computation, as we have seen that the changes in the prior are negligible.

We performed a simulation study of the frequentist performance of our prior and the one proposed by BBS. We ran 20,000 simulations of sample size $m = 1$ for values of $p$ equal to 0.05, 0.25, 0.50, 0.75, and 0.95. Similar to Section 5, we have considered the frequentist coverage of the 95% credible intervals and the frequentist relative MSE from the posterior means. Figure 7 shows the results for $p = 0.50$. We note very similar performances for both priors, with a slightly better MSE result for our prior. For the other values of $p$ (not shown here), the outcome is similar, with a more marked better performance of our prior for lower values of $p$ than for higher ones.

We have not examined in detail the case for $p$ unknown, as it has been done in BBS. However, the task does not bear any complication, as we can define the prior on $n$ and $p$ and $\pi(n, p) \propto \pi(n \mid p) \pi(p)$, where $\pi(p) \sim Be(1/2, 1/2)$ (Jeffreys’ prior) would be a natural choice. That is,

$$\begin{align*}
\pi(n, p) &\propto \frac{1}{(n+1)(1-p)} \exp \left\{ \sum_{x=0}^{n} \log(n+1-x) \binom{n}{x} \right\} \\
&\times p^{x}(1-p)^{n-x} - 1 \times \frac{1}{\sqrt{p(1-p)}},
\end{align*}$$

which, even though improper, will yield a proper posterior when the sample size if larger than or equal to 1.
7. DISCUSSION

The strength in deriving prior distributions by being objective, not by defining probabilities directly, but in assigning a worth to each element of the parameter space, is that it can be applied to any discrete parameter. It has been shown that such an objective prior can be defined for all the models presented in this article, and that these prior distributions are suitable for inference.

We believe the reason why this approach works can be found in the solid foundations it lies upon. Bayes’ rule, when applied in an objective setup, carries an important inconsistency, which is represented by the fact that \( \pi(\theta) \) (the prior) and \( \pi(\theta \mid x) \) (the posterior) do not carry the same interpretation. By resetting the update rule in the context of losses or utilities, the inconsistency disappears.

This approach can be extended to any situation where a prior has to be defined for a discrete parameter. A further noteworthy application can be found in Villa and Walker (2014), where a prior distribution for the number of degrees of freedom of a \( \ell \) distribution is derived.

The frequentist analysis of the models has shown that, in general, our prior performs at least as well as the BBS prior. The comparisons have been made by using very small sample sizes (i.e., \( m = 1 \)) where the possible differences in the frequentist properties of the priors are more prominent. For realistic values of sample size both priors appear to have very similar frequentist properties.

For the practitioner, we would argue that a solid explanation of the prior, subjective or objective, is mandatory. This becomes a more acute obligation when the prior is objective. Our prior has such an explanation in terms of measuring the worth of a parameter value being included in the model. As such a user can find a comprehensible interpretation of the prior and fully understand the derivation of \( \pi(\theta) \) for each \( \theta \). Moreover, the approach to derive our prior is applicable to any discrete parameter model.

Finally, by considering the limiting process in Blyth (1994), the idea we propose can be extended to continuous parameters. That is,

\[
\lim_{\delta \to 0} \frac{1}{\delta^2} D_{\text{KL}}(f(x \mid \theta) \parallel f(x \mid \theta + \delta)) = \sum_{i,j} I_{i,j}(\theta),
\]

where \( I_{i,j}(\theta) \) is the \((i, j)\)th element of the Fisher information matrix. In Brown and Walker (2012), it is shown that the approach, in the continuous case, may yield Jeffreys’ prior.

SUPPLEMENTARY MATERIAL

Proof of Lemmas and Theorems: This file contains the proofs to lemmas and theorems of the main article.

[Received March 2013. Revised May 2014.]

REFERENCES

Barger, K., and Bunge, J. (2008), “Bayesian Estimation of the Number of Species Using Noninformative Priors,” Biometrical Journal, 50, 1064–1076. [1072]

Berger, J. O., and Bernardo, J. M. (1992), “On the Development of the Reference Priors,” (with discussion), in Bayesian Statistics 4, eds. J. M. Bernardo, J. O. Berger, D. V. Lindley, and A. F. M. Smith, London: Oxford University Press, pp. 35–60 [1072]

Berger, J. O., Bernardo, J. M., and Sun, D. (2009), “The Formal Definition of Reference Priors,” The Annals of Statistics, 37, 905–938. [1072]

——— (2012), “Objective Priors for Discrete Parameter Spaces,” Journal of the American Statistical Association, 107, 636–648. [1072]

Berk, R. H. (1966), “Limiting Behaviour of Posterior Distributions When the Model is Incorrect,” Annals of Mathematical Statistics, 37, 51–58. [1073]

Bernardo, J. M. (1979), “Reference Posterior Distributions for Bayesian Inference” (with discussion), Journal of the Royal Statistical Society, Series B, 41, 113–147. [1072]

——— (2005), “Reference Analysis,” in Handbook of Statistics (Vol. 25), eds. D. K. Dey and C. R. Rao., Amsterdam: Elsevier, pp. 17–90 [1072]

Bernardo, J. M., and Smith, A. F. M. (1994), Bayesian Theory, New York: Wiley. [1072,1074]

Blyth, S. (1994), “Local Divergence and Association,” Biometrika, 81, 579–584. [1082]

Brown, P. J., and Walker, S. G. (2012), “Bayesian Priors from Loss Matching” (with discussion), International Statistical Review, 80, 60–82. [1082]

Goudie, I. B. J., and Goldie, C. M. (1981), “Initial Size Estimation for the Pure Death Process,” Biometrika, 68, 543–550. [1073]

Jeffreys, H. (1961), Theory of Probability, London: Oxford University Press. [1076]

Kullback, S., and Leibler, R. A. (1951), “On Information and Sufficiency,” Annals of Mathematical Statistics, 22, 79–86. [1073]

Lindsay, B. G., and Roeder, K. (1987), “A Unified Treatment of Integer Parameter Models,” Journal of the American Statistical Association, 82, 758–764. [1072]

Merhav, N., and Feder, M. (1998), “Universal Prediction,” IEEE Transactions on Information Theory, 44, 2124–2147. [1073]

Rissanen, J. (1983), “A Universal Prior for Integers and Estimation by Minimum Description Length,” The Annals of Statistics, 11, 416–431. [1072]

Villa, C., and Walker, S. G. (2014), “Objective Prior for the Number of Degrees of Freedom of a \( \ell \) Distribution,” Bayesian Analysis, 9, 197–220. [1082]