entropy

MDPI

Article

Prior Elicitation, Assessment and Inference with a Dirichlet Prior

Michael Evans *, Irwin Guttman and Peiying Li

Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada; sttirwin@buffalo.edu (I.G.); peiying.li@mail.utoronto.ca (P.L.)

* Correspondence: mevans@utstat.utoronto.ca; Tel.: +1-416-287-7274

Received: 29 July 2017; Accepted: 20 October 2017; Published: 22 October 2017

Abstract: Methods are developed for eliciting a Dirichlet prior based upon stating bounds on the individual probabilities that hold with high prior probability. This approach to selecting a prior is applied to a contingency table problem where it is demonstrated how to assess the prior with respect to the bias it induces as well as how to check for prior-data conflict. It is shown that the assessment of a hypothesis via relative belief can easily take into account what it means for the falsity of the hypothesis to correspond to a difference of practical importance and provide evidence in favor of a hypothesis.

Keywords: elicitation; bias; prior-data conflict; relative belief inferences; multinomial distribution; Dirichlet prior

1. Introduction

Perhaps the most basic statistical model is the multinomial $(n, p_1, \ldots, p_k)$ where $n \in \aleph$, $(p_1, \ldots, p_k) \in S_k = \{(x_1, \ldots, x_k) : x_i \geq 0 \text{ and } x_1 + \cdots + x_k = 1\}$, S_k is the $(k-1)$-dimensional simplex and $(p_1, \ldots, p_k)$ is unknown. This arises from an i.i.d. sample from the multinomial$(1, p_1, \ldots, p_k)$ distribution. The goal is then inference about the unknown value of $(p_1, \ldots, p_k)$.

Bayesian inference requires a prior and the Dirichlet$(\alpha_1, \ldots, \alpha_k)$, for some choice of hyperparameters $\alpha_1, \ldots, \alpha_k \geq 0$, is a convenient choice due to its conjugacy. The prior density is of the form $\pi(p_1, \ldots, p_k) = d(\alpha_1, \ldots, \alpha_k) p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1}$ where $d(\alpha_1, \ldots, \alpha_k) = \Gamma(\alpha_1 + \cdots + \alpha_k)/\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)$ for $(p_1, \ldots, p_k) \in S_k$. To employ such a prior it is necessary to have an elicitation algorithm to determine the hyperparameters. The purpose of this paper is to develop a particular algorithm based on bounds on the probabilities; to show how the chosen prior can be assessed with respect to the bias that it induces; to demonstrate how to check whether or not the prior conflicts with the data; to show how to modify the prior when such a conflict is encountered; and to implement inferences using the prior based on a measure of statistical evidence.

A key component of a Bayesian statistical analysis is the choice of the prior. This paper is concerned with the choice of a proper prior for a statistical analysis. It is generally acknowledged that the correct way to do this is through a process of elicitation where knowledgeable experts translate what is known about an application into the choice of a probability distribution reflecting beliefs about the unknown values of certain quantities. This is in contrast to the use of rules for the choice of default priors which are supposedly objective, such as the use of the principle of insufficient reason or the use of a Jeffreys prior. In fact, such default priors are also subjectively chosen as there appears to be no universal rule for this purpose and the specific rule itself needs to be chosen. In addition, these rules sometimes produce priors with characteristics that imply very specific beliefs, such as the Jeffreys prior for the multinomial which is a Dirichlet with all hyperparameters equal to 1/2. In essence, elicitation is honest about the subjectivity inherent in the choice of the prior and provides an argument for why the choice was made. In the context of the Dirichlet this knowledge will take the form of how likely a success is expected on each of the k categories being counted. Discussions about the process of elicitation for general problems can be found in [1,2].

In Section 2, current methods for eliciting a Dirichlet prior are reviewed and a new method is developed that possesses some advantages for situations where a weakly informative prior is required. Perhaps the main difference between the elicitation algorithm developed here and those already available in the literature, is that the user is required to state a lower or upper bound on each probability that they are virtually certain holds. Thus, a user knows that a cell probability must be smaller than some upper bound or knows that a probability must be larger than some lower bound. Rather than stating that such bounds hold categorically, the bound is believed to hold with a large prior probability, hence the terminology "virtual certainty". This follows good practice as the support of the prior is still the whole simplex and so does not rule out any values as being impossible. Note that the lower bound of 0 and the upper bound of 1 on a probability always hold with absolute certainty, so there is no concern that such bounds cannot be provided, but in many cases much tighter bounds will be applicable. One of the primary contributions of this paper is show how these bounds can be chosen consistently in the sense that they determine a Dirichlet prior and to develop an algorithm for obtaining this prior. In addition, it is shown in an example that this approach lends itself very naturally to determining a prior for the testing of independence. It is to be noted, however, that no elicitation methodology can be viewed as the correct approach and the existence of many approaches can only help to encourage the broad and effective use of priors. Thus, for a particular problem another elicitation algorithm, such as one among those reviewed in Section 2, may be felt to be more suitable.

A prior chosen via elicitation is proper. This allows for criticism of the prior in light of the observed data, namely, an assessment for prior-data conflict. If a prior is found to be in conflict with the data then, unless there is so much data that the effect of the prior is negligible, it is necessary to modify the prior to avoid this. These issues are discussed in Section 3.2.

In addition one has to be concerned about whether or not the choice of the prior results in bias. In fact, the issue of bias could be considered one of the main reasons for doubts being expressed about the appropriateness of Bayesian methodology. To precisely define bias it seems necessary to formulate a measure of evidence and here we use the relative belief ratio which is the ratio of the posterior to the prior as this measures change in belief from a priori to a posteriori. The assessment of bias in the prior, using this measure of evidence, is addressed in Section 3.1.

All inferences are derived from the relative belief ratio. Such inferences are invariant under 1-1 increasing functions of the relative belief ratio (as well as being invariant under smooth reparameterizations) and so the measure of evidence can equivalently be defined as the log of the relative belief ratio. It is then immediate that the expected evidence under the posterior is the relative entropy between the posterior and prior. In essence the relative entropy is a kind of average evidence and the log of the relative belief ratio at a specific parameter value is playing the role of the bit in the definition of entropy. It is to be noted, however, that for inference the concern is with measuring evidence, either in favor of or against a specific value, and not with the measurement of the more abstract concept of information. As such, there is an intimate connection between the concepts of entropy, evidence and relative belief inferences. Our purpose here, however, is not to consider this connection but discuss a methodology for choosing a prior for one of the most basic statistical problems, demonstrate how the chosen prior is to be assessed for conflict with data and for bias and then used for the derivation of inferences. Relative belief inferences for the multinomial are discussed in Section 4.

This presents a full treatment of a statistical analysis for the multinomial, although it is assumed that the multinomial model is correct. Strictly speaking, provided the data are available, it should also be checked that the initial sample is i.i.d. from a multinomial$(1, p_1, \ldots, p_k)$ distribution, perhaps using a multivariate version of a runs test, but this is not addressed here.

Throughout the paper, Π denotes the prior probability measure on the full model parameter θ, which in the case of the multinomial is the vector of cell probabilities, and π denotes its density.

Dependence of Π on hyperparameters is indicated by subscripts, such as $\Pi_{(\alpha_1,\ldots,\alpha_k)}$ denoting the Dirichlet$(\alpha_1, \ldots, \alpha_k)$ distribution. When a particular prior Π is referenced and interest is in the marginal prior of some function $\psi = \Psi(\theta)$, then Π_Ψ is used for the marginal prior measure of ψ with corresponding density π_Ψ. In addition, M denotes the prior (predictive) probability measure of the data induced by Π and the sampling model and m denotes the corresponding density.

The following example, taken from [3], is considered as a practical application of the methodology.

Example 1. *Assessing independence.*

Individuals were classified according to their blood type Y (O, A, B, and AB, although the AB individuals were eliminated, as they were small in number) and also classified according to X, their disease status (peptic ulcer = P, gastric cancer = G, or control = C). Thus, there are three populations; namely, those suffering from a peptic ulcer, those suffering from gastric cancer, and those suffering from neither, and it is assumed that the individuals involved in the study can be considered as random samples from the respective populations. The data are in Table 1 and the goal is to determine whether or not X and Y are independent. Thus, the counts are assumed to be multinomial$(8766, p_{11}, p_{12}, p_{13}, p_{21}, p_{22}, p_{23}, p_{31}, p_{32}, p_{33})$ *where the first index refers to X and the second to Y and with a relabelling of the categories, e.g., $X = G$ is relabeled as $X = 2$.*

Table 1. The data in Example 1.

	$Y = O$	$Y = A$	$Y = B$	**Total**
$X = P$	983	679	134	1796
$X = G$	383	416	84	883
$X = C$	2892	2625	570	6087
Total	4258	3720	788	8766

Using the chi-squared test, the null hypothesis of no relationship is rejected with a value of the chi-squared statistic of 40.54 *and a p-value of* 0.0000. *Table 2 gives the estimated cell probabilities based on the full multinomial as well as the estimated cell probabilities based on independence between X and Y. The difference between the two tables is very small and of questionable practical significance. For example, the largest difference between corresponding cells is* 0.012 *and, as a natural measure of difference between two distributions, the estimated Kullback-Leibler divergence, based on the raw data, is estimated as* 0.002. *This suggests that in reality the deviation from independence is not meaningful. The cure for this is that, in assessing any hypothesis, it is necessary to say what size of deviation δ from the null is of practical significance and take this into account when performing the test. This arises as a natural aspect of the relative belief approach to this problem and will be discussed in Sections 3 and 4 it is shown that a very different conclusion is reached in this example.*

Table 2. The estimated cell probabilities in Example 1 based on the full and independence models.

Full	$Y = O$	$Y = A$	$Y = B$	**Ind.**	$Y = O$	$Y = A$	$Y = B$
$X = P$	0.112	0.077	0.015	$X = P$	0.100	0.087	0.018
$X = G$	0.043	0.047	0.009	$X = G$	0.049	0.043	0.009
$X = C$	0.330	0.299	0.065	$X = C$	0.337	0.295	0.062

2. Elicitation

The problem of eliciting a Dirichlet prior is simplest when $k = 2$ and this corresponds to a beta distribution. Since this simple case contains the essence of the approach to elicitation for the Dirichlet presented here, this is considered first.

2.1. Eliciting a Beta Prior

Consider first the situation where $k = 2$ and the prior Π_{α_1,α_2} on p_1 is beta(α_1, α_2). Suppose it is known with "virtual certainty" that $l_1 \le p_1 \le u_1$ where $l_1, u_1 \in [0, 1]$ are known. This immediately implies that $1 - u_1 \le p_2 = 1 - p_1 \le 1 - l_1$ with virtual certainty. Here "virtual certainty" is interpreted to mean that the true value of p_1 is in the interval $[l_1, u_1]$ with high prior probability γ, say $\gamma = 0.99$. Thus, this restricts the prior to those values of (α_1, α_2) satisfying $\Pi_{\alpha_1,\alpha_2}([l_1, u_1]) = \gamma$. Note that in general there may be several values of (α_1, α_2) that satisfy this equality. For example, if $l_1 = 1/2 - a$ and $u_1 = 1/2 + a$ with $a > 0$, then $\Pi_{\alpha,1}([l_1, u_1]) = \Pi_{1,\alpha}([l_1, u_1])$ for all α. To completely determine (α_1, α_2) another condition is added, namely, it is required that the mode of the prior be at the point $\xi \in [l_1, u_1]$ as this allows the placement of the primary amount of the prior mass at an appropriate place within $[l_1, u_1]$. For example, a natural choice of the mode in this context is $\xi = (l_1 + u_1)/2$, namely, the midpoint of the interval. When $\alpha_1, \alpha_2 \ge 1$ the mode of the beta(α_1, α_2) occurs at $\xi = (\alpha_1 - 1)/\tau$ where $\tau = \alpha_1 + \alpha_2 - 2$. There is thus a 1-1 correspondence between the values (α_1, α_2) and (ξ, τ) given by $\alpha_1 = 1 + \tau\xi, \alpha_2 = 1 + \tau(1 - \xi)$. Hereafter, we restrict to the case $\alpha_i \ge 1$ to avoid singularities on the boundary as these seem difficult to justify a priori. Therefore, after specifying the mode, only the scaling of the beta prior is required through the choice of τ. Now if $X \sim$beta$(1 + \tau\xi, 1 + \tau(1 - \xi))$, then $E(X) = (1 + \tau\xi)/(2 + \tau) \to \xi$ and $Var(X) = (1 + \tau\xi)(1 + \tau(1 - \xi))/(2 + \tau)^2(3 + \tau) \to 0$, as $\tau \to \infty$, which establishes that $\Pi_{1+\tau\xi,1+\tau(1-\xi)}([l_1, u_1]) \to 1$ as $\tau \to \infty$. Thus, the following result has been proven since $\Pi_{1,1}([l_1, u_1]) = u_1 - l_1$.

Theorem 1. *For $0 \le l_1 < u_1 \le 1, \gamma \in (0, 1)$ and $\xi \in [l_1, u_1]$, then the beta$(1 + \tau\xi, 1 + \tau(1 - \xi))$ distribution has its mode at ξ and whenever $u_1 - l_1 \le \gamma$, there is a value $\tau \in [0, \infty)$ such that there is exactly γ of the probability in $[l_1, u_1]$.*

While the theorem establishes the existence of a value τ satisfying the requisite equation, it does not establish that this value is unique. Although uniqueness is not necessary for the methodology, based on examples and intuition, it seems very likely that $\Pi_{1+\tau\xi,1+\tau(1-\xi)}([l_1, u_1])$ is a monotone increasing function of τ which would imply that the τ in Theorem 1 is in fact unique. In any case, τ can be computed by choosing $\tau_0 = 0$, finding a value τ_* such that $\Pi_{1+\tau_*\xi,1+\tau_*(1-\xi)}([l_1, u_1]) > \gamma$ and then obtaining $\tau \in [\tau_0, \tau_*]$ satisfying the equality via the bisection root finding algorithm. This procedure is guaranteed to converge by the intermediate value theorem.

Example 2. *Determining a beta prior.*

Suppose that $[l_1, u_1] = (0.25, 0.75), \xi = 0.5$ and $\gamma = 0.99$. The solution obtained via the iterative algorithm is then $\tau = 22.0$ where the iteration is stopped when $|\Pi_{1+\tau_i\xi,1+\tau_i(1-\xi)}([l_1, u_1]) - \gamma| \le 0.005$. This took seven iterations, the prior is given by $(\alpha_1, \alpha_2) = (12.0, 12.0)$ and $[l_1, u_1]$ contains 0.993 of the prior probability. If instead of 0.005 the error tolerance for stopping was set equal to 0.001, then the solution $\tau = 22.04$ and $(\alpha_1, \alpha_2) = (12.02, 12.02)$ was obtained after 20 iterations with $[l_1, u_1]$ containing 0.990 of the prior probability.

If $u_1 - l_1 > \gamma$, then $\Pi_{1,1}([l_1, u_1]) > \gamma$ and virtual certainty for $[l_1, u_1]$ is obtained by $(\alpha_1, \alpha_2) = (1, 1)$.

The concept of "virtual certainty" is interpreted as something being true "with high probability" and choosing γ close to 1 reflects this. For example, in rolling an apparently symmetrical die the analyst may be quite certain that the probability p_i of observing i pips is a least $1/8$ and wants the prior to reflect this. In effect, the goal is to ensure that the prior concentrates its mass in the region satisfying these inequalities and choosing γ large accomplishes this. Actually, it is not necessary that exact equality is obtained to ensure virtual certainty. As long as γ is close to 1, then small changes in γ will not lead to big changes in the prior as in Example 2 where it is seen that choosing $\gamma = 0.993$ rather than $\gamma = 0.990$ makes very little difference in the prior. Specifying probabilities beyond 2 or 3 decimal places seems impractical in most applications so taking γ in the range $[0.990, 0.999]$ seems quite satisfactory for characterizing virtual certainty while allowing some flexibility for the analyst. Far more important than the choice of γ is the selection of what it is felt is known, for example, the bounds l_1 and u_1 on the probabilities for the beta prior, as mistakes can be made. Protection against a misleading analysis caused by a poor choice of a prior is approached through checking for prior-data conflict and modifying the prior appropriately when this is the case, as discussed in Section 3. It is also to be noted that the methodology does not require γ to be large as the analyst may only be willing to say that the bounds on the probabilities for the die hold with prior probability $\gamma = 0.50$. However, choosing the bounds so that these are fairly weak constraints on the probabilities, and so almost certainly hold as is reflected by choosing γ close to 1, seems like an easy way to be weakly informative.

2.2. *Eliciting a Dirichlet Prior*

The approach to eliciting a beta prior allows for a great deal of flexibility as to where the prior allocates the bulk of its mass in $[0, 1]$. The question, however, is how to generalize this to the Dirichlet$(\alpha_1, \ldots, \alpha_k)$ prior. As will be seen, it is necessary to be careful about how $(\alpha_1, \ldots, \alpha_k)$ is elicited. Again, we make the restriction that each $\alpha_i \geq 1$ to avoid singularities for the prior on the boundary.

It seems quite natural to think about putting probabilistic bounds on the p_i, such as requiring $l_i \leq p_i \leq u_i$ with high probability, for fixed constants l_i, u_i, to reflect what is known with virtual certainty about p_i. For example, it may be known that p_i is very small and so we put $l_i = 0$, choose u_i small and require that $p_i \leq u_i$ with prior probability at least γ. While placing bounds like this on the p_i seems reasonable, such an approach can result in a complicated shape for the region that is to contain the true value of $(p_1, \ldots, p_k)$ with virtual certainty. This complexity can make the computations associated with inference very difficult. In fact, it can be hard to determine exactly what the full region is. As such, it seems better to use an elicitation method that fits well with the geometry of the Dirichlet family. If it is felt that more is known a priori than a Dirichlet prior can express, then it is appropriate to contemplate using some other family of priors, see, for example, Elfadaly and Garthwaite [4,5]. Given the conjugacy property of Dirichlet priors and their common usage, the focus here is on devising elicitation algorithms that work well with this family. First, however, we consider elicitation approaches for this problem that have been presented in the literature.

Chaloner and Duncan [6] discuss an iterative elicitation algorithm based on specifying characteristics of the prior predictive distribution of the data which is Dirichlet-multinomial. Regazzini and Sazonov [7] discuss an elicitation algorithm which entails partitioning the simplex, prescribing prior probabilities for each element of the partition and then selecting a mixture of Dirichlet distributions such that this prior has Prohorov distance less than some $\epsilon > 0$ from the true prior associated with de Finetti's representation theorem. Both of these approaches are complicated to implement. Closest to the method presented here is that discussed in [8] where $(\alpha_1, \ldots, \alpha_k)$ is specified by choosing $i \in \{1, \ldots, k\}$, stating two prior quantiles $(p_{\gamma_{i1}}, p_{\gamma_{i2}})$ where $0 < \gamma_{i1} < \gamma_{i2} < 1$ for p_i and specifying prior quantile p_{γ_j} for p_j for each $j \neq i, k$. Thus, there are k constraints that the Dirichlet$(\alpha_1, \ldots, \alpha_k)$ has to satisfy and an algorithm is provided for computing $(\alpha_1, \ldots, \alpha_k)$. Drawbacks include the fact that the p_i are not treated symmetrically as there is a need to place two constraints on one of the probabilities and p_k is treated quite differently than the other probabilities. In addition, precise quantiles need to be specified and values $\alpha_i < 1$ can be obtained which induce singularities in the prior. Furthermore, it is not at all clear what these constraints say about the joint prior on $(p_1, \ldots, p_k)$ as this elicitation does not take into account the dependencies that occur necessarily among the p_i. Zapata-Vázquez et al. [9] develop an elicitation algorithm based on eliciting beta distributions for the individual probabilities and then constructing a Dirichlet prior that represents a compromise among these marginals. Elfadaly and Garthwaite [4] determine a Dirichlet by eliciting the first quartile, median and third quartile for the conditional distribution of $p_i \mid p_1, \ldots, p_{i-1}$ and finding the beta distribution, rescaled by the factor $1 - \sum_{j=1}^{i-1} p_j$, that best fits these quantiles. This requires the prescription of precise quantiles, an order in which to elicit the conditionals and an iterative approach to reconcile the elicited conditional quantiles when these quantiles are not consistent with a Dirichlet. A notable aspect of their approach is that it also works for the Connor-Mosimann distribution, a generalization of the Dirichlet, and in that case no reconciliation is required. Similarly, Elfadaly and Garthwaite [5] base the elicitation on the three quartiles of the marginal beta distributions of the p_i which, while independent of order, still requires reconciliation to ensure that the elicited marginals correspond to a Dirichlet. In addition, the elicitation procedure based on the conditionals is extended to develop an elicitation procedure for a more flexible prior based on a Gaussian copula.

The approach in this paper is based on the idea of placing bounds on the probabilities that hold with virtual certainty and that are mutually consistent for any prior on S_k. The user need only check that the bounds stated satisfy the conditions stated in the theorems to ensure consistency and these can be very simple to check and modify appropriately. Rather than being required to state precise quantiles or moments for the prior, all that is required are weak bounds on the probabilities. For example, we might be willing to say that we are virtually certain that p_i is greater than a value l_i. We consider l_i a weak bound because there may be some belief that the true value is much greater than l_i but being precise about how to express such beliefs is more difficult and requires more refined judgements. Certainly elicitation methodology that requires more assessment than what is being required here is even more open to concerns about robustness and other issues with the prior. As discussed in Sections 3–5, such concerns are better addressed through considerations about prior-data conflict, bias and using inference methods that are as robust to the prior as possible.

There are several versions depending on whether lower or upper bounds are placed on the p_i. We start with the situation where a lower bound is given for each p_i as this provides the basic idea for the others. Generally the elicitation process allows for a single lower or upper bound to be specified for each p_i. These bounds specify a subsimplex of the simplex S_k with all edges of the same length. As will be seen, this implicitly takes into account the dependencies among the p_i. With such a region determined, it is straightforward to find $(\alpha_1, \ldots, \alpha_k)$ such that the subsimplex contains γ of the prior probability for $(p_1, \ldots, p_k)$. It is worth noting that the bounds determined in Theorems 2–4 can be applied to any family of priors on S_k and it is only in Section 2.2.4 where specific reference is made to the Dirichlet.

Note that a $(k-1)$-simplex can be specified by k distinct points in R^k, say $\mathbf{a}_1, \ldots, \mathbf{a}_k$, and then taking all convex combinations of these points. This simplex will be denoted as $S(\mathbf{a}_1, \ldots, \mathbf{a}_k) = \{\sum_{i=1}^k c_i \mathbf{a}_i : c_i \geq 0 \text{ with } c_1 + \cdots + c_k = 1\}$. Thus, $S_k = S(\mathbf{e}_1, \ldots, \mathbf{e}_k)$, where $\mathbf{e}_i$ is the i-th standard basis vector of R^k, and it is clear that $S(\mathbf{a}_1, \ldots, \mathbf{a}_k) \subset S_k$ whenever $\mathbf{a}_1, \ldots, \mathbf{a}_k \in S_k$. The *centroid* of $S(\mathbf{a}_1, \ldots, \mathbf{a}_k)$ is equal to $CS(\mathbf{a}_1, \ldots, \mathbf{a}_k) = \sum_{i=1}^k \mathbf{a}_i / k$.

2.2.1. Lower Bounds on the Probabilities

For this we ask for a set of lower bounds $l_1, \ldots, l_k \in [0,1]$ such that $l_i \leq p_i$ for $i = 1, \ldots, k$. To make sense, there is only one additional constraint that the l_i must satisfy, namely, $L_{1:k} = l_1 + \cdots + l_k \leq 1$. If $L_{1:k} = 1$, then it is immediate that $p_i = l_i$, otherwise $p_1 + \cdots + p_k > 1$. Thus, the p_i are completely determined when $L_{1:k} = 1$. Attention is thus restricted to the case where $L_{1:k} < 1$. The following result then holds.

Theorem 2. *Specifying the lower bounds $l_1, \ldots, l_k \in [0,1]$ such that $l_i \leq p_i$ for $i = 1, \ldots, k$ and*

$$L_{1:k} < 1, \tag{1}$$

prescribes $S(\mathbf{a}_1, \ldots, \mathbf{a}_k) \subset S_k$ where $\mathbf{a}_i = (l_1, \ldots, l_{i-1}, u_i, l_{i+1}, \ldots, l_k)$ and

$$u_i = 1 - \sum_{j \neq i} l_j. \tag{2}$$

The edges of $S(\mathbf{a}_1,\ldots,\mathbf{a}_k)$ *each have length* $\sqrt{2}(1-L_{1:k})$ *and* $S(\mathbf{a}_1,\ldots,\mathbf{a}_k)=\{(p_1,\ldots,p_k):p_1+\cdots+p_k=1,l_i\le p_i\le u_i,i=1,\ldots,k\}$.

Proof. Note that (1) implies that $p_i=1-\sum_{j\neq i}p_j\le 1-\sum_{j\neq i}l_j=u_i$, and so stating the lower bounds implies a set of upper bounds, and also $l_i<u_i\le 1$. Consider now the set $S=\{(p_1,\ldots,p_k): p_1+\cdots+p_k=1,l_i\le p_i\le u_i,i=1,\ldots,k\}$ and note that $\mathbf{a}_i\in S$ for $i=1,\ldots,k$. For $c_i\ge 0$ with $c_1+\cdots+c_k=1$, then $(p_1,\ldots,p_k)=\sum_{i=1}^k c_i\mathbf{a}_i\in S$ since, for example, the first coordinate satisfies $p_1=c_1u_1+(\sum_{i=2}^k c_i)l_1=c_1u_1+(1-c_1)l_1$ so $l_1\le p_1\le u_1$. Therefore $S(\mathbf{a}_1,\ldots,\mathbf{a}_k)\subset S$.

If $(p_1,\ldots,p_k)\in S$, then $p_i=c_i^*l_i+(1-c_i^*)u_i$ where $c_i^*\in[0,1]$. Now $1=p_1+\cdots+p_k=\sum_{i=1}^k c_i^*l_i+\sum_{i=1}^k(1-c_i^*)u_i=\sum_{i=1}^k c_i^*l_i+\sum_{i=1}^k(1-c_i^*)\,(l_i+1-L_{1:k})=L_{1:k}+\{\sum_{i=1}^k(1-c_i^*)\}\,(1-L_{1:k})$ and so $\sum_{i=1}^k(1-c_i^*)=1$. For $(p_1,\ldots,p_k)=\sum_{j=1}^k(1-c_j^*)\mathbf{a}_j$ we have $p_i=(\sum_{j\neq i}(1-c_j^*))l_i+(1-c_i^*)u_i=c_i^*l_i+(1-c_i^*)u_i$. This proves that $S\subset S(\mathbf{a}_1,\ldots,\mathbf{a}_k)$ and so we have $S(\mathbf{a}_1,\ldots,\mathbf{a}_k)=S$.

Finally note that $||\mathbf{a}_i-\mathbf{a}_j||^2=(u_i-l_i)^2+(u_i-l_j)^2=2(1-L_{1:k})^2$ and so $S(\mathbf{a}_1,\ldots,\mathbf{a}_k)$ has edges all of the same length. This completes the proof. □

It is relatively straightforward to ensure that the elicited bounds are consistent with a prior on S_k. For, if it is determined that $L_{1:k}\ge 1$, then it is simply a matter of lowering some of the bounds to ensure (1) is satisfied. For example, multiplying all the bounds by a common factor can do this and lowered l_i means greater conservatism as it is a weaker bound. Furthermore, it is perfectly acceptable to set some $l_i=0$ as this does not affect the result.

2.2.2. Upper Bounds on the Probabilities

Of course, it may be that prior beliefs are instead expressed via upper bounds on the probabilities or a mixture of upper and lower bounds. The case of all upper bounds is considered first. Our goal is to specify the upper bounds in such a way that these lead unambiguously to lower bounds $l_1,\ldots,l_k\in[0,1]$ satisfying (1) and so to the simplex $S(\mathbf{a}_1,\ldots,\mathbf{a}_k)$.

Suppose then that we have the upper bounds $u_1,\ldots,u_k\in[0,1]$ such that $p_i\le u_i$. It is clear then that $l_1,\ldots,l_k$ must satisfy the system of linear equations given by (2) as well as $0\le l_i\le u_i$ for $i=1,\ldots,k$ and (1). Thus, the l_i must satisfy

$$\mathbf{u}=\mathbf{1}_k-\begin{pmatrix}0&1&\ldots&1\\1&0&\ldots&1\\\vdots&\vdots&\vdots&\vdots\\1&1&\ldots&0\end{pmatrix}\boldsymbol{l}=\mathbf{1}_k+(I_k-\mathbf{1}_k\mathbf{1}_k')\boldsymbol{l} \tag{3}$$

where $\mathbf{1}_k$ is the k-dimensional vector of 1's and I_k is the $k\times k$ identity. Noting that $(I_k-\mathbf{1}_k\mathbf{1}_k')^{-1}=I_k-(k-1)^{-1}\mathbf{1}_k\mathbf{1}_k'$, it is immediate that

$$\boldsymbol{l}=(I_k-(k-1)^{-1}\mathbf{1}_k\mathbf{1}_k')(\mathbf{u}-\mathbf{1}_k). \tag{4}$$

Note that this requires that $k\ge 2$ as is always the case.

Putting $U_{1:k}=\sum_{j=1}^k u_j$, then (4) implies $L_{1:k}=(k-U_{1:k})/(k-1)$ and so $0\le L_{1:k}<1$ provided $U_{1:k}$ satisfies

$$1<U_{1:k}\le k. \tag{5}$$

From (4)

$$l_i=(u_i-1)-\frac{U_{1:k}-k}{k-1}=u_i+\frac{1-U_{1:k}}{k-1} \tag{6}$$

and, for $i=1,\ldots,k$, this implies that $l_i\ge 0$ iff

$$u_i\ge\frac{U_{1:k}-1}{k-1}. \tag{7}$$

In addition, when (5) is satisfied, then $l_i < u_i$ for $i = 1, \ldots, k$. This completes the proof of the following result.

Theorem 3. *Specifying upper bounds $u_1, \ldots, u_k \in [0, 1]$, such that $p_i \leq u_i$ for $i = 1, \ldots, k$, satisfying inequalities (5) and (7), determines the lower bounds $l_1, \ldots, l_k$, given by (6), which determine the simplex $S(\mathbf{a}_1, \ldots, \mathbf{a}_k)$ defined in Theorem 2.*

For this elicitation to be consistent with a prior on S_k it is necessary to make sure that the upper bounds satisfy (5) and (7). If we take $u_1 = \cdots = u_k = u \geq 1/k$, then (5) is satisfied and $(k - 1)u \geq ku - 1$ implies that (7) is satisfied as well. If $U_{1:k} \leq 1$, then the u_i need to be increased which is conservative and note that $U_{1:k} \leq k$ is true provided all the $u_i \leq 1$ which is always the case. If (5) is satisfied but (7) is not for some i, then u_i must be increased, which is again conservative, and (5) is still satisfied. Thus, again, making sure the elicited bounds are consistent is straight-forward. In addition, the bound $u_i = 1$ is an acceptable choice.

2.2.3. Upper and Lower Bounds on the Probabilities

Now, perhaps after relabelling the probabilities, suppose that lower bounds $0 \leq l_i \leq p_i$ for $i = 1, \ldots, m$ as well as upper bounds $p_i \leq u_i \leq 1$ for $i = m + 1, \ldots, k$, where $1 \leq m < k$, have been provided. Again, it is required that $L_{1:m} = l_1 + \cdots + l_m < 1$ and we search for conditions on the u_i that complete the prescription of a full set of lower bounds $l_1, \ldots, l_k$ so that Theorem 2 applies. Again the $\mathbf{l}$ and $\mathbf{u}$ vectors must satisfy (3). Let $\mathbf{x}_{r:s}$ denote the subvector of $\mathbf{x}$ given by its consecutive r-th through s-th coordinates and $X_{r:s}$ the sum of these coordinates provided $r \leq s$ and be null otherwise. The following equations hold

$$\begin{aligned} \mathbf{u}_{1:m} &= \mathbf{1}_m + \boldsymbol{l}_{1:m} - L_{1:m}\mathbf{1}_m - L_{m+1:k}\mathbf{1}_m \\ \mathbf{u}_{m+1:k} &= \mathbf{1}_{k-m} - L_{1:m}\mathbf{1}_{k-m} + (I_{k-m} - \mathbf{1}_{k-m}\mathbf{1}'_{k-m})\boldsymbol{l}_{m+1:k}. \end{aligned}$$

Rearranging these equations so the knowns are on the left and the unknowns are on the right gives

$$\boldsymbol{l}_{1:m} + (1 - L_{1:m})\mathbf{1}_m = \mathbf{u}_{1:m} + L_{m+1:k}\mathbf{1}_m \tag{8}$$

$$\mathbf{u}_{m+1:k} - (1 - L_{1:m})\mathbf{1}_{k-m} = (I_{k-m} - \mathbf{1}_{k-m}\mathbf{1}'_{k-m})\boldsymbol{l}_{m+1:k}. \tag{9}$$

It follows from (9) that

$$\begin{aligned} \boldsymbol{l}_{m+1:k} &= (I_{k-m} - \mathbf{1}_{k-m}\mathbf{1}'_{k-m})^{-1}[\mathbf{u}_{m+1:k} - (1 - L_{1:m})\boldsymbol{l}_{k-m}] \\ &= (I_{k-m} - (k - m - 1)^{-1}\mathbf{1}_{k-m}\mathbf{1}'_{k-m})[\mathbf{u}_{m+1:k} - (1 - L_{1:m})\boldsymbol{l}_{k-m}] \end{aligned} \tag{10}$$

and substituting this into (8) gives the solution for $\mathbf{u}_{1:m}$ as well.

Thus, it is only necessary to determine what additional conditions have to be imposed on the $l_1, \ldots, l_m, u_m, \ldots, u_k$ so that Theorem 2 applies. Note that it follows from (8) that $\mathbf{u}_{1:m}$ takes the correct form, as given by (2), so it is really only necessary to check that $\boldsymbol{l}$ is appropriate.

First it is noted that it is necessary that $k - m > 1$. The case $k - m = 1$ only occurs when $m = k - 1$ and then $p_k = 1 - p_1 - \cdots - p_{k-1} \leq 1 - l_1 - \cdots - l_{k-1}$ which is the required value for u_k for Theorem 2 to apply. Thus, when $k - m = 1$, there is no choice but to put $u_k = 1 - l_1 - \cdots - l_{k-1}$ and choose a lower bound for p_k, which of course could be 0, which means that Theorem 2 applies. It is assumed hereafter that $k - m > 1$.

Now $L_{1:k} = L_{1:m} + L_{m+1:k}$ and the requirement $0 \leq L_{1:k} < 1$ imposes the requirement $0 \leq L_{m+1:k} < 1 - L_{1:m}$. Using (10) gives

$$\begin{aligned}L_{m+1:k} = \mathbf{1}'_{k-m}\boldsymbol{l}_{m+1:k} &= \left(1 - \frac{k-m}{k-m-1}\right)(U_{m+1:k} - (k-m)(1-L_{1:m})) \\ &= \frac{(k-m)(1-L_{1:m}) - U_{m+1:k}}{k-m-1}\end{aligned}$$

and therefore $0 \le L_{m+1:k} < 1 - L_{1:m}$ iff

$$1 - L_{1:m} < U_{m+1:k} \le (k-m)(1-L_{1:m}). \tag{11}$$

It is seen that (11) generalizes (5) on taking $m = 0$. Now for $i > m$

$$\begin{aligned}l_i &= u_i - (1-L_{1:m}) - \frac{U_{m+1:k}}{k-m-1} + \frac{(k-m)(1-L_{1:m})}{k-m-1} \\ &= u_i + \frac{(1-L_{1:m}) - U_{m+1:k}}{k-m-1}\end{aligned} \tag{12}$$

thus, for $i = m+1, \ldots, k$, this implies that $l_i \ge 0$ iff

$$u_i \ge \frac{U_{m+1:k} - (1-L_{1:m})}{k-m-1}. \tag{13}$$

Thus, (13) generalizes (5) on taking $m = 0$. In addition, if (11) is satisfied, then $l_i \le u_i$ for $i = m+1, \ldots, k$.

The above argument establishes the following result.

Theorem 4. *For m satisfying $1 \le m \le k-2$, specifying the bounds*

(i) $l_i \le p_i$ *with* $l_i \in [0,1]$ *for* $i = 1, \ldots, m$, *satisfying* $L_{1:m} < 1$; *and*
(ii) $u_i \ge p_i$ *with* $u_i \in [0,1]$ *for* $i = m+1, \ldots, k$, *satisfying (11) and (13), determines the lower bounds* $l_{m+1}, \ldots, l_k$, *given by (12), which, together with* $l_1, \ldots, l_m$, *determine the simplex* $S(\mathbf{a}_1, \ldots, \mathbf{a}_k)$ *defined in Theorem 2.*

Ensuring that the elicited bounds are consistent with a prior on S_k can proceed as follows. First ensuring $L_{1:m} < 1$ can be accomplished conservatively by lowering some of the l_i if necessary. In addition, the inequality $1 - L_{1:m} < U_{m+1:k}$ can be accomplished conservatively by raising some of the u_i if necessary. If $U_{m+1:k} > (k-m)(1-L_{1:m})$, then some of the u_i need to be decreased or some of the l_i need to be increased or a combination of both. Indeed setting a $u_i = 1$ to be conservative, so (13) is satisfied, may require lowering some of the lower bounds but again this is conservative. Note that, if we assign the u_i such that $U_{m+1:k} = (k-m)(1-L_{1:m})$, then (13) reduces to $u_i \ge 1 - L_{1:m}$ and the assignment $u_i = 1 - L_{1:m}$ ensures consistency although an alternative assignment can be made such that $U_{m+1:k} = (k-m)(1-L_{1:m})$ holds.

The purpose of Theorems 2–4 is to ensure that the bounds selected for the individual probabilities are consistent. It may be that an expert has a bound which they believe holds with virtual certainty but the consistency requirements are violated. The solution to this problem is to decrease a lower bound or increase an upper bound so that the requirements are satisfied. While this is not an entirely satisfactory solution to this problem, it does not violate the prescription that the bounds hold with virtual certainty. Furthermore, the lower bound of 0, or the upper bound of 1, is always available if a user feels they have absolutely no idea how to choose such a bound.

2.2.4. Determining the Elicited Dirichlet Prior

Theorems 2–4 state bounds that are consistent for a prior on S_k. Thus, now it is necessary to determine the Dirichlet$(\alpha_1, \ldots, \alpha_k)$ prior, denoted $\Pi_{(\alpha_1,\ldots,\alpha_k)}$, such that $\Pi_{(\alpha_1,\ldots,\alpha_k)}(S(\mathbf{a}_1, \ldots, \mathbf{a}_k)) = \gamma$. Again we pick a point $\xi = (\xi_1, \ldots, \xi_k) \in S(\mathbf{a}_1, \ldots, \mathbf{a}_k)$ and place the mode at ξ, so $\xi_i = (\alpha_i - 1)/\tau$ for $i = 1, \ldots, k$ with $\tau = \alpha_1 + \cdots + \alpha_k - k$. For example, $\xi = CS(\mathbf{a}_1, \ldots, \mathbf{a}_k)$ would often seem like a sensible choice and then only τ needs to be determined. There is a 1-1 correspondence between $(\alpha_1, \ldots, \alpha_k)$ and $(\xi_1, \ldots, \xi_k, \tau)$ given by $\alpha_i = 1 + \tau\xi_i$.

Again it makes sense to proceed via an iterative algorithm to determine τ. Provided $\Pi_{(1,\ldots,1)}(S(\mathbf{a}_1, \ldots, \mathbf{a}_k)) \leq \gamma$, set $\tau_0 = 0$ and find τ_1 such that $\Pi_{(1+\tau_i\xi_1,\ldots,1+\tau_i\xi_k)}(S(\mathbf{a}_1, \ldots, \mathbf{a}_k)) \geq \gamma$. As before set $\tau_2 = (\tau_1 + \tau_0)/2$ and then the algorithm proceeds via bisection. Determining $\Pi_{(1+\tau_i\xi_1,\ldots,1+\tau_i\xi_k)}(S(\mathbf{a}_1, \ldots, \mathbf{a}_k))$ at each step becomes problematical even for $k = 3$. In the approach adopted here this probability content was estimated via a Monte Carlo sample from the relevant Dirichlet. This is seen to work quite well as, in the case of determining a prior, high accuracy for the computations is not required.

Consider an example.

Example 3. *Determining a Dirichlet*$(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ *prior.*

Suppose that $k = 4$ *and the lower bounds* $l_1 = 0.2, l_2 = 0.2, l_3 = 0.3, l_4 = 0.2$ *are placed on the probabilities. This results in the bounds* $0.2 \leq p_1 \leq 0.3, 0.2 \leq p_2 \leq 0.3, 0.3 \leq p_3 \leq 0.4,$ *and* $0.2 \leq p_4 \leq 0.3$ *which are reasonably tight. The mode was placed at the centroid* $\xi = (0.22, 0.22, 0.32, 0.22)$. *For* $\gamma = 0.99$, *an error tolerance of* $\epsilon = 0.005$ *and a Monte Carlo sample of size of* $N = 10^3$ *at each step, the values* $\tau = 2560$ *and* $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (577.0, 577.0, 833.0, 577.0)$ *were obtained after 13 iterations. The prior content of* $S(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, , \mathbf{a}_4)$ *was estimated to be* 0.989. *If greater accuracy is required then N can be increased and/or* ϵ *decreased.*

This choice of lower bounds results in a fairly concentrated prior as is reflected in the plots of the marginals in Figure 1. This concentration is not a defect of the elicitation as (2) indicates that it must occur when the sum of the bounds is close to 1. Thus, the concentration is forced by the dependencies among the probabilities.

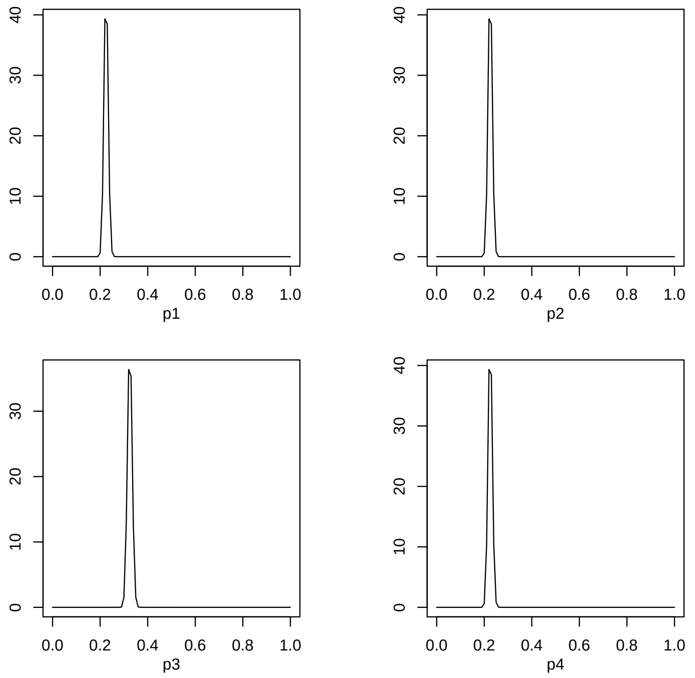

Figure 1. Plots of the marginal densities determined when specifying the lower bounds $l_1 = 0.2$, $l_2 = 0.2, l_3 = 0.3, l_4 = 0.2$ in Example 3.

Consider now another example.

Example 4. *Determining a Dirichlet* $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9)$ *prior.*

Suppose that $k = 9$ *and the lower bounds* $l_1 = 0.02, l_2 = 0.02, l_3 = 0.0, l_4 = 0.00, l_5 = 0.00$, $l_6 = 0.00$, $l_7 = 0.10, l_8 = 0.10, , l_9 = 0.00$ *are placed on the probabilities. This leads to the following bounds for the probabilities.*

$$\begin{array}{lll} 0.02 \le p_1 \le 0.78 & 0.02 \le p_2 \le 0.78 & 0.00 \le p_3 \le 0.76 \\ 0.00 \le p_4 \le 0.76 & 0.00 \le p_5 \le 0.76 & 0.00 \le p_6 \le 0.76 \\ 0.10 \le p_7 \le 0.86 & 0.10 \le p_8 \le 0.86 & 0.00 \le p_9 \le 0.76 \end{array}$$

The mode was placed at the centroid $\xi = (0.1, 0.1, 0.08, 0.08, 0.08, 0.08, 0.18, 0.18, 0.08)$*. For* $\gamma = 0.99$, *an error tolerance of* $\epsilon = 0.005$ *and a Monte Carlo sample of size of* $N = 10^3$ *at each step, the values* $\tau = 96$ *and* $(\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7, \alpha_8, \alpha_9) = (11.03, 11.03, 9.11, 9.11, 9.11, 9.11, 18.71, 18.71, 9.11)$ *were obtained after seven iterations. The prior content of* $S(\mathbf{a}_1, \ldots, \mathbf{a}_9)$ *was estimated to be* 0.987*. Figure 2 is a plot of the nine marginal priors for the* p_i*. Again, the dependencies among the* p_i *make the marginal priors quite concentrated.*

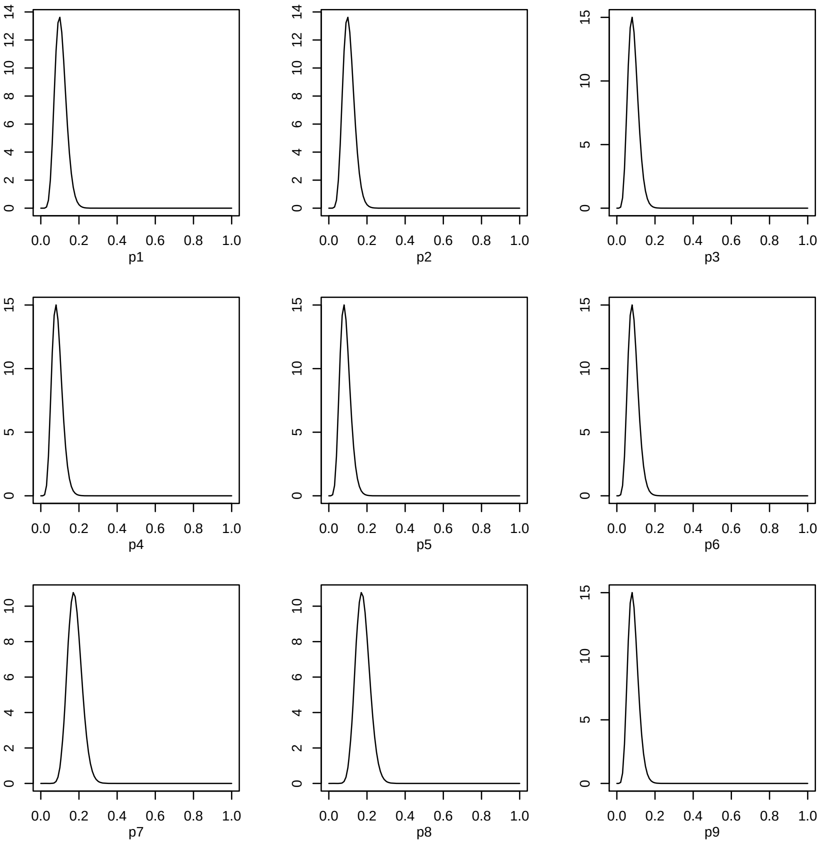

Figure 2. Plot of the nine marginal priors in Example 4.

Example 1 (continued). *Choosing the prior.*

Given that we wish to assess independence, it is necessary that any elicited prior include independence as a possibility so this is not ruled out a priori. A natural elicitation is to specify valid bounds (namely, bounds that satisfy our theorems) on the $p_{i\cdot}$ *and the* $p_{\cdot j}$ *and then use these to obtain bounds on the* p_{ij} *which in turn leads to the prior. Thus, suppose valid bounds have been specified that lead to the lower bounds* $a_i \le p_{i\cdot}, b_j \le p_{\cdot j}$*. Then it is necessary that* $l_{ij} = a_i b_j$ *is the lower bound on* p_{ij}*. Note that it is immediate that the* l_{ij} *satisfy the conditions of Theorem 2 and from (2),* $p_{ij} \le 1 - \sum_{r,s} l_{rs} + l_{ij} = 1 - \sum_r a_r \sum_s b_s + a_i b_j$ *which is greater than* $l_{ij} = a_i b_j$ *since* $0 \le \sum_r a_r < 1$ *and* $0 \le \sum_s b_s < 1$*. As such the region for the* p_{ij} *contains elements of* H_0.

For this example, the lower bounds $a_1 = 0.1, a_2 = 0.0, a_3 = 0.5, b_1 = 0.2, b_2 = 0.2, b_3 = 0.0$ *were chosen which leads to the lower bounds*

$$L = \begin{pmatrix} 0.02 & 0.02 & 0.00 \\ 0.00 & 0.00 & 0.00 \\ 0.10 & 0.10 & 0.00 \end{pmatrix}$$

on the p_{ij}*. Note that these are precisely the bounds used in Example 4 so the prior is as determined in that example where the indexing is row-wise.*

The software used in this paper to determine the prior from the bounds is available at http://utstat.utoronto.ca/mikevans/software/Dirichlet/RDirichlet.html.

3. Assessing the Prior

Here, we specialize the developments discussed in [10,11] to the multinomial problem with a Dirichlet prior. It is to be noted that the methods presented in this section for the assessment of a prior are applicable to any prior and not just in the special circumstances discussed here.

Suppose a quantity $\psi = \Psi(p_1, \ldots, p_k)$ is of interest and there is a need to assess the hypothesis $H_0 : \Psi(p_1, \ldots, p_k) = \psi_0$. Let π_Ψ denote the prior density and $\pi_\Psi(\cdot \,|\, f_1, \ldots, f_k)$ denote the posterior density of Ψ, where $(f_1, \ldots, f_k)$ gives the observed cell counts. When $\Psi(p_1, \ldots, p_k) = (p_1, \ldots, p_k)$, then π_Ψ is the Dirichlet$(\alpha_1, \ldots, \alpha_k)$ density and $\pi_\Psi(\cdot \,|\, f_1, \ldots, f_k)$ is the Dirichlet$(\alpha_1 + f_1, \ldots, \alpha_k + f_k)$ density. The relative belief ratio $RB_\Psi(\psi_0 \,|\, f_1, \ldots, f_k)$ is defined as the limiting ratio of the posterior probability of a set containing ψ_0 to the prior probability of this set where the limit is taken as the set converges (nicely) to the point ψ_0. Whenever $\pi_\Psi(\psi_0) > 0$ and π_Ψ is continuous at ψ_0, then $RB_\Psi(\psi_0 \,|\, f_1, \ldots, f_k) = \pi_\Psi(\psi_0 \,|\, f_1, \ldots, f_k)/\pi_\Psi(\psi_0)$. As such, $RB_\Psi(\psi_0 \,|\, f_1, \ldots, f_k)$ is measuring how beliefs about ψ_0 have changed from a priori to a posteriori and is a measure of evidence concerning H_0. If $RB_\Psi(\psi_0 \,|\, f_1, \ldots, f_k) > 1$, then there is evidence that H_0 is true, as belief in the truth of H_0 has increased, if $RB_\Psi(\psi_0 \,|\, f_1, \ldots, f_k) < 1$, then there is evidence that H_0 is false, as belief in the truth of H_0 has decreased and if $RB_\Psi(\psi_0 \,|\, f_1, \ldots, f_k) = 1$, then there is no evidence either way.

Any 1-1 increasing transformation of a relative belief ratio can also be used to measure evidence. For example, $\log RB_\Psi(\psi_0 \,|\, f_1, \ldots, f_k)$ works just as well but now $\log RB_\Psi(\psi_0 \,|\, f_1, \ldots, f_k) > (<)\, 0$ provides evidence for (against) H_0. As mentioned in the Introduction, this establishes a connection between relative belief and relative entropy. The Bayes factor is the ratio of the posterior odds to prior odds and so is also a measure of change in belief and, as such, is a measure of evidence. When the prior on ψ is discrete, the Bayes factor for the event $\{\psi_0\}$ equals $RB_\Psi(\psi_0 \,|\, f_1, \ldots, f_k)/RB_\Psi(\{\psi_0\}^c \,|\, f_1, \ldots, f_k)$ where $\{\psi_0\}^c$ is the complement of $\{\psi_0\}$. Thus, the Bayes factor can be expressed in terms of the relative belief ratio but not conversely. Furthermore, it can be proved that when $RB_\Psi(\psi_0 \,|\, f_1, \ldots, f_k) > (<)\, 1$, then $RB_\Psi(\{\psi_0\}^c \,|\, f_1, \ldots, f_k) < (>)\, 1$ which simply expresses the natural property that evidence for $\{\psi_0\}$ is evidence against $\{\psi_0\}^c$ and conversely. Thus, it is seen that the relative belief ratio is a more fundamental measure of evidence and moreover the Bayes factor is not really comparing the evidence for $\{\psi_0\}$ with the evidence for its negation. When the prior on ψ is continuous, the issue is more complicated because of a common recommendation that such a prior be replaced by a mixture with a point mass at ψ_0 so that a Bayes factor can be defined. Alternatively, one could define the Bayes factor at ψ_0 in the continuous case as the limit of Bayes factors of shrinking sets as we have done for the relative belief ratio. When this definition is used, the Bayes factor is identical to the relative belief ratio. For these reasons, and a number of optimality properties proven for relative belief ratios, we adopt the relative belief ratio as the basic measure of evidence. These issues and results are more fully discussed in [11].

3.1. Assessing Bias in the Prior

Given that there is a measure of evidence for H_0, it is possible to assess the bias in the prior with respect to H_0. For this let $M(\cdot \,|\, \psi_0)$ denote the prior predictive distribution of $(f_1, \ldots, f_k)$ given that $\Psi(p_1, \ldots, p_k) = \psi_0$. The bias against H_0 is assessed by

$$M(RB_{\Psi}(\psi_0 \,|\, f_1, \ldots, f_k) \leq 1 \,|\, \psi_0), \tag{14}$$

the prior probability that evidence in favor of H_0 will not be obtained when H_0 is true. If (14) is large, then there is bias in the prior against H_0 and, as such, if evidence against H_0 is obtained after seeing the data, then this should have little impact. In essence the ingredients of the study are such that it is not meaningful to find evidence against H_0. To measure bias in favor of H_0, let ψ_* be a value of Ψ that is just meaningfully different than ψ_0. In other words, values ψ that differ from ψ_0 less than ψ_* does, are not considered as practically different than ψ_0. Then the bias in favor of H_0 is measured by

$$M(RB_{\Psi}(\psi_0 \,|\, f_1, \ldots, f_k) \geq 1 \,|\, \psi_*). \tag{15}$$

If (15) is large, then there is bias in favor of H_0 and if evidence in favor of H_0.is obtained after seeing the data, then this should have little impact. It is shown in [11] that both (14) and (15) converge to 0 as $n \to \infty$. Thus, bias can be controlled by sample size.

The computation of (14) and (15) can be difficult in certain contexts with the primary issue being the need to generate from the conditional prior predictives of the data. As in the following example, however, great accuracy is typically not required for these computations and so effective methods are available.

Example 1 (continued). *Measuring bias and choosing δ.*

To assess independence between X and Y, the marginal parameter

$$\psi = \Psi(p_{11}, p_{12}, \ldots, p_{kl}) = \sum_{i,j} p_{ij} \ln(p_{ij}/p_{i\cdot}p_{\cdot j}) \tag{16}$$

is used. Note that (16) is the minimum Kullback-Leibler distance between the p_{ij} values and an element of H_0. Furthermore, $\psi = 0$ iff independence holds.

As discussed previously, it is necessary to specify a $\delta > 0$ such that a practically meaningful lack of independence occurs iff the true value $\psi \geq \delta$. One approach is to specify a δ such that, if $-\delta \leq (p_{ij} - p_{i\cdot}p_{\cdot j})/p_{ij} < \delta$ for all i and j, then any such deviation is practically insignificant, as the relative errors are all bounded by δ. Using $\ln(1+x) \approx x$ for small x, this condition implies that $-\delta \leq \psi < \delta$. The range of ψ is then discretized using this δ and the hypothesis to be assessed is now, because $\psi \geq 0$ always, $H_0 : 0 \leq \psi < \delta$. This assessment is carried out using the relative belief ratios based on the discretized prior and posterior of Ψ as discussed in Section 4. For the data in this problem, we take $\delta = 0.01$ which corresponds to a 1% relative error. Thus, this says that we do not consider independence as failing when the true probabilities differ from probabilities based on independence with a relative error of less than 1%.

With this choice of δ the issue of bias is now addressed. The prior distribution of the discretized Ψ is determined by simulation. For this, generate the p_{ij} from the elicited prior and compute ψ and the prior probability contents of the intervals for ψ given by $[0, \delta), [\delta, 2\delta), \ldots, [(k-1)\delta, k\delta)$ where k is determined so as to cover the full range of observed generated values of ψ. The plot of the prior density histogram for ψ is provided in Figure 3.

For inference, the posterior contents of these intervals are also determined via simulating from the posterior based on the observed data. For measuring bias, however, we proceed as follows. Each time a generated ψ satisfies $[0, \delta)$ the corresponding p_{ij} are used to generate a new data set F_{ij} and $RB_{\Psi}([0, \delta) \,|\, F_{11}, \ldots, F_{kl})$ is determined and note that this requires generating from the posterior based on the F_{ij}. The probability

$M(RB_{\Psi}([0,\delta)\,|\,F_{11},\ldots,F_{kl}) \le 1\,|\,[0,\delta))$ is then estimated by the proportion of these relative belief ratios that are less than or equal to 1. This gives an estimate of the bias against H_0. Estimating the bias in favor of H_0 proceeds similarly, but now the F_{ij} are generated whenever $\psi \in [\delta, 2\delta)$ is satisfied, as these represent values that correspond to just differing from independence meaningfully.

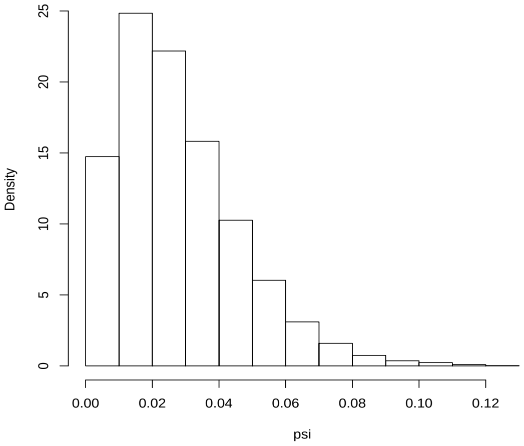

Figure 3. Plot of the prior density histogram for ψ in Example 1.

Clearly this procedure could be computationally quite demanding if highly accurate estimates of the biases are required. In general, however, high accuracy is not necessary. Even accuracy to one decimal place will provide a clear indication of whether or not there is serious bias. In this problem the biases for the elicited prior are estimated to be 0.12 *for bias for and* 0.02 *for bias against. Thus, there is only a probably of* 0.02 *of obtaining evidence against H_0 when it is true, which implies virtually no bias against H_0. There is, however, a prior probability of* 0.12 *of obtaining evidence in favor of H_0 when it is just meaningfully false and so some bias in favor of H_0. It is to be noted that bias decreases as ψ_* moves away from ψ_0. These values depend on the chosen value of δ but in fact are reasonably robust to this choice. The prior probability content of the interval* $[0, 0.01)$ *is* 0.14 *while* $[0.01, 0.02)$ *contains* 0.25 *of the prior probability. Thus, there is a reasonable amount of prior probability allocated to effective independence and also to the smallest nonindependence of interest.*

3.2. Checking for Prior-Data Conflict

Anytime a prior is used it is reasonable to question whether or not the prior is contradicted by the data. Essentially such a contradiction occurs when the data indicate that the true value of the model parameter lies in the tails of the prior. While opinions vary on this, the point-of-view taken here is that properly collected data are primary in determining inferences, and so models and priors that are contradicted by the data need to be modified when this occurs. The issue is somewhat less relevant for priors, as with enough data the effect of the prior is minimal, but on the other hand it often turns out to be relatively easy to modify the prior so that the conflict is avoided, see [12].

The elicitation discussed here could be in error, namely, if the true probabilities lie well outside the intervals obtained. If the data demonstrate this in a reasonably conclusive way, then it would seem incorrect to proceed with an analysis based on this prior unless there was an absolute conviction that the amount of data was sufficient to overwhelm the influence of the prior. To check for prior-data conflict we follow Evans and Moshonov [13] and compute the tail probability

$$M(m(F_1,\ldots,F_k) \le m(f_1,\ldots,f_k)) \tag{17}$$

where $(f_1,\ldots,f_k)$ is the observed value of the minimal sufficient statistic and M is the prior predictive distribution of this statistic with density m. In [14] it is proved that quite generally (17) converges to

$\Pi(\pi(p_1,\ldots,p_k) \le \pi(p_{1,true},\ldots,p_{k,true}))$ as $n \to \infty$, where Π is the prior on $(p_1,\ldots,p_k)$. Thus, a small value of (17) is indicating that the true value of $(p_1,\ldots,p_k)$ lies in a region where the prior is relatively low and so the data are contradicting the prior. Certainly a value like 0.01 for (17) suggests that the true value is well into the "tails" of the prior. It is to be noted that prior-data conflict can have a number of ill-effects. For example, results in [15] show that robustness to the prior cannot be achieved in the presence of prior-data conflict.

When the prior is given by the uniform, then a simple computation shows that (17) is equal to 1 and so there is no prior-data conflict. Intuitively, the closer τ is to 0, then the less information the prior is putting into the analysis. This idea can be made precise in terms of the weak informativity of one prior with respect to another as developed in [12]. As such, if prior-data conflict is obtained with the prior specified by a value of $(\xi_1,\ldots,\xi_k,\tau)$, then this prior can be replaced by a prior that is weakly informative with respect to it so that the conflict can be avoided and this entails choosing a value $\tau' < \tau$.

Example 1 (continued). *Checking the elicited prior.*

For the elicited Dirichlet prior, the value of (17) is approximately equal to 1 (to the accuracy of the computations) and so there is definitely no prior-data conflict.

4. Inference

For data $(f_1,\ldots,f_k)$ and Dirichlet$(\alpha_1,\ldots,\alpha_k)$ prior the posterior, of $(p_1,\ldots,p_k)$ is Dirichlet$(\alpha_1 + f_1,\ldots,\alpha_k + f_k)$. As such it is easy to generate from the posterior of ψ, estimate the posterior contents of the intervals $[(i-1)\delta, i\delta)$ and then estimate the relative belief ratios $RB_\Psi([(i-1)\delta, i\delta) \,|\, f_1,\ldots,f_k)$. From this a relative belief estimate of the discretized ψ can be obtained and various hypotheses assessed for this quantity.

The strength of the evidence provided by $RB_\Psi(\psi_0 \,|\, f_1,\ldots,f_k)$ is measured by, see [11],

$$\Pi_\Psi(RB_\Psi(\psi \,|\, f_1,\ldots,f_k) \le RB_\Psi(\psi_0 \,|\, f_1,\ldots,f_k) \,|\, f_1,\ldots,f_k), \tag{18}$$

namely, the posterior probability that the true value of ψ has a relative belief ratio no greater than the hypothesized value. When $RB_\Psi(\psi_0 \,|\, f_1,\ldots,f_k) < 1$, so there is evidence against ψ_0, a small value for (18) implies there is strong evidence against ψ_0 since there is a large posterior probability that the true value has a larger relative belief ratio than ψ_0. When $RB_\Psi(\psi_0 \,|\, f_1,\ldots,f_k) > 1$, so there is evidence in favor of ψ_0, a large value for (18) indicates there is strong evidence in favor of ψ_0 since there is a small posterior probability that the true value has a larger relative belief ratio than ψ_0. Note that when $RB_\Psi(\psi_0 \,|\, f_1,\ldots,f_k) > 1$, then the best estimate of ψ in the set $\{\psi : RB_\Psi(\psi \,|\, f_1,\ldots,f_k) \le RB_\Psi(\psi_0 \,|\, f_1,\ldots,f_k)\}$ is ψ_0 as it has the most evidence in its favor. While the measure of strength looks like a p-value, it has a very different interpretation and it is not measuring evidence. Note that, if our goal was instead to estimate ψ, then the measure of evidence adopted dictates that this be given by the relative belief estimate $\psi(x) = \arg\sup_\psi RB_\Psi(\psi \,|\, f_1,\ldots,f_k)$ as this is the value with the most evidence in its favor ($\sup_\psi RB_\Psi(\psi \,|\, f_1,\ldots,f_k)$ is always greater than 1). In addition, an assessment of the accuracy of the estimate is given by the size of a λ-relative belief region $C_\lambda(f_1,\ldots,f_k) = \{\psi : RB_\Psi(\psi \,|\, f_1,\ldots,f_k) \ge c_\lambda(f_1,\ldots,f_k)\}$ where $c_\lambda(f_1,\ldots,f_k)$ is the smallest constant so that the posterior content of $C_\lambda(f_1,\ldots,f_k)$ is a least λ for a choice of $\lambda \in (0,1)$. Note that $\psi(x)$ is always in $C_\lambda(f_1,\ldots,f_k)$. Relative belief inferences possess a number of optimal properties in the class of Bayesian inferences, see [11], and with particular relevance for the choice of prior, optimal robustness to the prior properties as developed in [15]. Whether or not the elicitation methodology itself assists in inducing such robustness is a matter for further investigation. Especially when the bounds are chosen to be quite diffuse, this seems plausible.

Given that there is no prior-data conflict with the elicited prior and little or no bias in this prior relative to the hypothesis H_0 of independence, we can proceed to inference in Example 1.

Example 1 (continued). *Inference.*

The posterior of the p_{ij} *is the Dirichlet*$(998.2, 694.2, 146.48, 395.48, 428.48, 96.48, 2918.1, 2651.1, 582.48)$ *distribution. For the hypothesis* H_0 *of independence between the variables, and using the discretized Kullback-Leibler divergence with* $\delta = 0.01$, *the value* $RB_{\Psi}([0, \delta) \,|\, f_1, \ldots, f_k) = 7.13$ *was obtained so there is evidence in favor of* H_0. *For the strength of this evidence the value of (18) equals* 1. *Thus, the evidence in favor of* H_0 *is of the maximum possible strength. Of course, this is due to the large sample size and the fact that the posterior distribution concentrates entirely in* $[0, \delta)$. *Note that this is a very different conclusion than that obtained by the p-value based on the chi-squared test.*

5. Conclusions

A very natural and easy to use method has been developed for eliciting Dirichlet priors based upon placing bounds on the individual probabilities that takes into account the dependencies among the probabilities. Of course, there may be more information available, such as upper and lower bounds on many of the probabilities. The price paid for this, however, is a much more complicated region where the bulk of the prior mass is located and even difficulties in determining what that region is, so this represents a problem for further work. It is also relevant to consider the individual bounds holding with possibly different prior probabilities but, as with considering both lower and upper bounds simultaneously for each probability, mathematical issues arise that make this a problem for further work.

While we view the approach to elicitation presented here as being fairly simple, it is certainly reasonable that other approaches are practically useful and preferable in certain situations. There is no doubt that the Dirichlet imposes what may be unnatural constraints for some situations and so more general families of priors are also needed for the multinomial. As such, extending our approach to more general families of priors is another problem of interest. In particular, Theorems 2–4 are relevant to any family of priors placed on S_k and so, provided sampling from such a prior is straightforward and there is a nice way to parameterize the prior as with the Dirichlet, then the approach of this paper can be implemented.

The application of the Dirichlet prior to an inference problem has also been illustrated using a measure of statistical evidence, the relative belief ratio, as a basis for the inferences. Given that a measure of evidence has been identified, it is possible to assess the bias in the prior before proceeding to inference. In addition, the prior has been checked to see if it is contradicted by the data. While the adequacy of the prior in light of the data can be assessed via the methods discussed in Section 3, there is also a need to measure how closely an elicited prior reflects an expert's judgements and suitable methodology needs to be developed for that problem.

Finally, it is seen that the assessment of a hypothesis can be different than that obtained by a standard p-value and, in particular, provide evidence in favor of a hypothesis. Of course, this is based on a well-known defect in p-values, namely, with a large enough sample, a failure of the hypothesis of no practical importance can be detected. The solution to this problem is to say what difference matters and use an approach that incorporates this. Relative belief inferences are seen to do this in a very natural way. The choice of δ is not arbitrary but is rather a fundamental characteristic of the application. When such a δ cannot be determined, it is not a failure of the inference methodology, but rather reflects a failure of the analyst to understand an aspect of the application that is necessary for a more refined analysis to take place.

Acknowledgments: The work of Michael Evans was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors thank the reviewers for many constructive and helpful comments.

Author Contributions: Each author contributed equally to this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Garthwaite, P.H.; Kadane, J.B.; O'Hagan, A. Statistical methods for eliciting probability distributions. *J. Am. Stat. Assoc.* **2005**, *100*, 680-700.
2. O'Hagan, A.; Buck, C.E.; Daneshkhah, A.; Eiser, J.R.; Garthwaite, P.H.; Jenkinson, D.J.; Oakley, J.E.; Rakow, T. *Uncertain Judgements: Eliciting Experts' Probabilities*; Wiley: Chichester, UK, 2006.
3. Snedecor, G.; Cochran, W. *Statistical Methods*, 6th ed.; Iowa State University Press: Iowa City, IA, USA, 1967.
4. Elfadaly, F.G.; Garthwaite, P.H. Eliciting Dirichlet and Connor-Mosimann Prior Distributions for Multinomial Models. *Test* **2013**, *22*, 628–646.
5. Elfadaly, F.G.; Garthwaite, P.H. Eliciting Dirichlet and Gaussian copula prior distributions for multinomial models. *Stat. Comput.* **2017**, *27*, 449–467.
6. Chaloner, K.; Duncan, G.T. Some properties of the Dirichlet multinomial distribution and its use in prior elicitation. *Commun. Stat. Theory Methods* **1987**, *16*, 511–523.
7. Regazzini, E.; Sazonov, V.V. Approximation of laws of multinomial parameters by mixtures of Dirichlet distributions with applications to Bayesian inference. *Acta Appl. Math.* **1999**, *58*, 247–264.
8. Van Dorp, J.R.; Mazzuchi, T.A. Parameter specification of the beta distribution and its Dirichlet extensions utilizing quantiles. In *Handbook of Beta Distributions and Its Applications*; Gupta, A.K., Nadarajah, S., Eds.; Marcel Dekker: New York, NY, USA, 2004.
9. Zapata-Vázquez, R.; O'Hagan, A.; Bastos, L. Eliciting expert judgements about a set of proportions. *J. Appl. Stat.* **2014**, *41*, 1919–1933.
10. Baskurt, Z.; Evans, M. Hypothesis assessment and inequalities for Bayes factors and relative belief ratios. *Bayesian Anal.* **2013** *8*, 569–590.
11. Evans, M. *Measuring Statistical Evidence Using Relative Belief*; Monographs on Statistics and Applied Probability 144; CRC Press: Boca Raton, FL, USA, 2015.
12. Evans, M.; Jang, G.-H. Weak informativity and the information in one prior relative to another. *Stat. Sci.* **2011**, *26*, 423–439.
13. Evans, M.; Moshonov, H. Checking for prior-data conflict. *Bayesian Anal.* **2006**, *1*, 893-914.
14. Evans, M.; Jang, G.-H. A limit result for the prior predictive applied to checking for prior-data conflict. *Stat. Proba. Lett.* **2011**, *81*, 1034–1038.
15. Al-Labadi, L.; Evans, M. Optimal robustness results for some Bayesian procedures and the relationship to prior-data conflict. *Bayesian Anal.* **2017** 12, 702-728.