Characterizing Dirichlet priors

Marcio A. Diniz [*], Jasper De Bock [†], Arthur Van Camp [†]

Abstract

The selection of prior distributions is a problem that has been heavily discussed since Bayes and Price published their paper in 1763. Conjugate priors became popular, largely because of their mathematical convenience. In this study, we justify the use of the conjugate combination of a Dirichlet prior and a multinomial likelihood by imposing a fundamental principle that we call partition invariance, alongside other requirements that are well known in the literature.

keywords: predictive inference, partition invariance, Johnson's sufficientness postulate, conjugate prior

1 Introduction

To perform Bayesian inference, the traditional first steps consist of selecting a statistical model—from which the likelihood is derived—and a prior probability distribution for the unknown parameters of this model.

[*] Universidade Federal de S. Carlos, Dept. of Statistics

[†] Ghent University, SYSTeMS Research Group

Since Bayes and Price (1763) published their essay, the matter of the selection of the prior has received considerable attention. Two major points of view emerged, and the debate continues. Subjectivists argue that a prior is a purely subjective assessment that should be elicited from experts (Savage 1954, O'Hagan et al. 2006), whereas objectivists argue for the use of a unique objective prior, that can be obtained by imposing strict rules that follow from the selected statistical model or from logical principles (Kass and Wasserman 1996, Berger, Bernardo and Sun 2009).

A third perspective is called the logical approach, which can be regarded as a compromise between the two approaches presented above. As in the objective approach, the idea is to impose structural criteria. However, these criteria are typically less restrictive than those used by objectivists and thus do not lead to a unique prior, but to a fairly broad—and often parameterized—class of priors. From a subjective perspective, this class can then be regarded as a set of candidate priors, among which a single prior can be selected subjectively. Alternatively, an objective prior can be selected from within this set. See (Carnap 1952) and (Jaynes 2003) for more information on the logical approach.

Our perspective can be regarded as a generalization of the third approach. However, instead of focusing only on the selection of the prior, we regard the selection of the statistical model as part of the problem. The main idea is to reduce the class of all probabilistic models—combinations of statistical models and priors—available to select from by requiring that the induced predictive distribution satisfies certain principles. Additionally, we do not assume that the predictive distribution results from the conjunction of a statistical model and a prior; we infer this fact from basic principles.

In this paper, we develop such an approach for a generalized version of the problem posed by Bayes and Price (1763), also referred to as the fundamental problem of practical statistics by Karl Pearson (1920): in a binary experiment, given that we have observed k successes in the first n trials, what is the probability that the next trial will be a success?[1] This question is easily generalized to the non-binary case and to any (finite) number of future trials.

To reduce the class of priors and statistical models for this problem, we require predictive distributions to fulfill philosophical (or practical) principles. Exchangeability and coherence are well known and widely discussed in the literature; see (Good 1965), (Zabell 2005), (Jaynes 2003), (Robert 2007) and references therein. The open-mindedness condition and learning from experience are considered reasonable properties of a predictive distribution as well; see (Skyrms 1996), (Carnap 1952) and (Carnap and Stegmüller 1959). We combine these properties with an additional requirement, which we call the partition invariance principle. This principle is closely related to W. E. Johnson's sufficientness postulate, also called predictive sufficiency. Eventually, we prove that when a subject is studying a phenomenon for which those principles are considered necessary, and decides that his predictive distribution should reflect these principles, then his predictive distribution should be derived from the multinomial likelihood and a prior on the parameter space that is Dirichlet. In this way, we justify the use of this conjugate—and thus tractable and mathematically convenient—combination, based on fundamental principles.

[1] This problem is closely related to the problem of induction: how to justify inductive inference from the observed to the unobserved. This problem was posed by Hume (1888) and concerns the justification of inductive methods, i.e., methods that assume that "instances of which we have had no experience resemble those of which we have had experience".

This paper is organized as follows. Section 2 starts from a prior and a multinomial likelihood, and shows how they induce a predictive distribution. In Section 3, we reverse the process. We start from a predictive distribution and, by requiring it to satisfy certain principles, show that this predictive distribution is induced by a prior and a multinomial likelihood. In the non-binary case, we justify the use of Dirichlet priors by additionally imposing Johnson's sufficientness postulate. Then, in Section 4, we define the partition invariance principle and show that if a predictive distribution system—a map from sample spaces to predictive distributions—satisfies this principle, then each of the predictive distributions of which it consists satisfies Johnson's sufficientness postulate. Ultimately, this result allows us to justify the use of Dirichlet priors, even in the binary case. We close with some remarks and directions for future research.

2 From priors to predictions

As noted in the introduction, we consider a generalized version of the fundamental problem of statistics. Formally, this problem is concerned with an experiment that can be repeated indefinitely and for which the outcome that is realized in every repetition takes values in some finite category set $\mathscr{X}$. We use $e_i \in \mathscr{X}$ to denote the outcome of experiment i. The goal is to determine a *predictive distribution*.

Definition 1. A predictive distribution $P_{\mathscr{X}}$ gives, for any finite initial sequence of realized experiments $\{X_i = e_i\}_{i=1}^{n}$, a probability for any future event—any logical proposition based on a finite number of future experiments.

Example 1 (running example). Consider a binary category set $\mathscr{X} = \{0,1\}$. If the first experiment resulted in $e_1 = 1$, then a predictive distribution $P_{\mathscr{X}}$ could for example predict that $P_{\mathscr{X}}(X_2 = 0 \mid e_1) = 1/3$ and $P_{\mathscr{X}}(X_2 = X_3 \mid e_1) = 2/3$. $\Diamond$

For $\mathscr{X} = \{0, 1\} = \{\text{‘failure’}, \text{‘success’}\}$, the—by now well-known—solution proposed by Bayes and Price (1763) was to consider a (potentially) infinite sequence of binary random quantities that are assumed to be conditionally independent and identically distributed given θ, the probability of success. Assuming a prior for θ, with distribution function $\Pi(\theta)$, the probability of any given sequence $(X_1 = e_1, \ldots, X_n = e_n) \in \{0, 1\}^n$, in which k out of n trials are successful, is then given by

$$\int_0^1 \theta^k (1-\theta)^{n-k} d\Pi(\theta) =: P_{\mathscr{X}}(X_1 = e_1, \ldots, X_n = e_n), \tag{1}$$

also known as the marginal distribution of the data (Robert 2007), the prior predictive distribution (Schervish 1995), or the *marginal predictive distribution.* Bayes and Price used a uniform prior, but—as we now all know—the same approach also works with other priors.

Running example. When the prior for θ is taken as a beta distribution with hyperparameters $\alpha > 0$ and $\beta > 0$, and when k successes occur in n experiments, it can be shown that

$$P_{\mathscr{X}}(X_1 = e_1, \ldots, X_n = e_n) = \frac{\Gamma(\alpha+\beta)\Gamma(k+\alpha)\Gamma(n-k+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)}, \tag{2}$$

which is known as the beta-binomial or Polya distribution. In the special case of the uniform prior with $\alpha = \beta = 1$, we find—for example—that: $P_{\mathscr{X}}(X_1 = 1) = 1/2$ and $P_{\mathscr{X}}(X_1 = 1, X_2 = 0) = 1/6$. $\Diamond$

This method for constructing a marginal predictive distribution is easily generalized to the multinomial case.

Notably, Definition 1 also covers the case of a null initial sequence of experiments, implying that the marginal predictive distribution is a part of $P_{\mathscr{X}}$. When at least one experiment is realized, the predictive distribution produces conditional probabilities similar to those in example 1. These conditional probabilities are often called conditional or posterior predictive probabilities (Bernardo and Smith 1994, Schervish 1995). We will refer to these probabilities as conditional predictive probabilities and will refer to the part of the predictive distribution that provides them as the *conditional predictive distribution*. In this way, the predictive distribution consists of two parts: the marginal predictive distribution and the conditional predictive distribution.

With this terminology in place, we can continue to develop a predictive distribution. Equation (1) already provides us with the marginal part. The next step is to use this marginal predictive distribution to derive the conditional part of the predictive distribution. The traditional approach is to apply Bayes's rule. However, in order to be able to do so, we must impose the following condition. Loosely speaking, it requires that any finite sequence of experiments is not considered impossible.

Definition 2. A predictive distribution is said to be *open-minded* when, according to the corresponding marginal predictive distribution, every finite sequence of experiments has a probability (strictly) greater than zero:

$$P_{\mathscr{X}}(X_1 = e_1, \ldots, X_n = e_n) > 0 \text{ for all } n < \infty \text{ and } (e_1, \ldots, e_n) \in \mathscr{X}^n. \tag{3}$$

See for example (Skyrms 1996). The same condition was referred to as regularity by Carnap (1952)—see (Carnap and Stegmüller 1959) as well—and also became known as *Cromwell's rule* after Lindley (1991).[2]

[2] In 1650, Cromwell, trying to convince the General Assembly of the Church of Scotland that their support for Charles II was misguided, said: "I beseech you, in the bowels of Christ, think

If we accept the open-mindedness condition, then providing a marginal predictive distribution is equivalent to providing a predictive distribution because, using Bayes's rule, all conditional predictive probabilities can be computed without worrying about probability zero and, together with the marginal predictive distribution, these conditional probabilities constitute the predictive distribution of Definition 1.

In some cases, it will be useful to focus on the 'immediate' part of the predictive distribution. This *immediate predictive distribution* consists of the immediate marginal predictive probabilities $P_{\mathscr{X}}(X_1 = e_1)$ and immediate conditional predictive probabilities $P_{\mathscr{X}}(X_{n+1} = e_{n+1} \mid \{X_i = e_i\}_{i=1}^n)$. Using the rules of probability calculus, the predictive distribution $P_{\mathscr{X}}$ can be derived from its immediate part. Consequently, there is a one-to-one correspondence between predictive distributions and immediate predictive distributions.

Running example. In our example, with $\mathscr{X} = \{0, 1\}$ and a beta prior with hyperparameters $\alpha = \beta = 1$, and for $e_1 = 1$, we find that

$$P_{\mathscr{X}}(X_2 = 0 \mid e_1) = \frac{P_{\mathscr{X}}(X_1 = 1, X_2 = 0)}{P_{\mathscr{X}}(X_1 = 1)} = \frac{1/6}{1/2} = \frac{1}{3}.$$

More generally, for any $\alpha > 0$ and $\beta > 0$, computing the conditional predictive probabilities is non-problematic because the open-mindedness condition is clearly satisfied. For example, when h out of m future trials are successful, we find that

$$P_{\mathscr{X}}(X_{n+1} = e_{n+1}, \ldots, X_{n+m} = e_{n+m} \mid \{X_i = e_i\}_{i=1}^n) = \frac{(\alpha + k)^{(h)}(n - k + \beta)^{(m-h)}}{(n + \alpha + \beta)^{(m)}}, \tag{4}$$

where $(\alpha + k)^{(h)} = (\alpha + k)(\alpha + k + 1) \ldots (\alpha + k + h - 1)$ is the rising factorial. In the particular case of the uniform prior with $\alpha = \beta = 1$ and for $m = h = 1$, we

it possible you may be mistaken." Lindley used this saying to illustrate the principle that no proposition, however implausible, should be assigned zero probability.

find that

$$P_{\mathscr{X}}(X_{n+1} = 1 \mid \{X_i = e_i\}_{i=1}^{n}) = {}^{(k+1)}/_{(n+2)},$$

which corresponds to what is perhaps the most famous immediate predictive distribution: Laplace's rule of succession. $\diamondsuit$

Laplace employed the approach described above, using uniform priors, to solve many practical problems. His justification for using the uniform prior is currently called the *principle of indifference*, also referred to as the *Bayes-Laplace postulate*, which is a basic symmetry argument.

As time progressed, other priors started being used and, in principle, one could use any possible distribution as a prior. Therefore, the obvious question—and by now arguably the most important question in Bayesian statistics—is how to select a prior.

One of the first criteria on which this selection was based was mathematical convenience. This criterion lead to the concept of conjugate priors, the use of which was disseminated by Raiffa and Schlaifer (1961) and, for the exponential family, justified on mathematical grounds by Diaconis and Ylvisaker (1979). The beta/Dirichlet prior is the conjugate prior for the binomial/multinomial likelihood. Since Good (1965), mixtures (convex combinations) of Dirichlets also became popular because they allow for a broader range of assessments and because they are relatively easy to compute with thanks to the conjugacy property.

As computational power increased, tractability became less of an issue, and Bayesian inference started to be applied in many fields. The question of how to select a prior, once again, received considerable attention.

Two main approaches emerged; see (Robert 2007, chap. 3) for an overview. Subjectivists argue that a prior must reflect scientific knowledge of experts and, as such, should be obtained through elicitation. See (O'Hagan et al. 2006) and (Seaman III et al. 2012); (Bedrick et al. 1996) provides a good example of the application of such an approach. In contrast, objectivists derive a prior by requiring that it satisfies some structural criterion; Kass and Wasserman (1996) provide a nice overview. Well-known examples are reference priors (Bernardo 1979 and Berger, Bernardo and Sun 2009) and approaches based on maximum entropy (Jaynes 2003) or invariance to reparameterizations (Jeffreys 1961).

We do not wish to enter this discussion here. Instead, as noted in the introduction, we will follow an approach that is similar to the logical one, as founded by the philosophers W. E. Johnson (1924) and Rudolf Carnap (1952); see Keynes (1921) and Jaynes (2003) as well. This approach did not become popular among statisticians, and is more known among science philosophers. The idea is to take a step back from the discussion and to argue for a broad—often parameterized—class of priors by imposing philosophical principles. In our case, this class of priors will be the conjugate class of Dirichlet priors. We do not discuss how to select a single prior from this class of Dirichlet priors. Subjective or objective arguments could be used for that purpose.

3 From predictions to priors

As explained in the previous section, a predictive distribution can be derived from a prior by combining the prior with a multinomial likelihood and applying Bayes's rule. We will now justify this approach by starting from a predictive distribution and showing that, when we require this predictive distribution to satisfy certain properties, there must indeed be a prior that induces it in this way. Eventually, by imposing additional principles, we will justify the use of Dirichlet priors.

A first important property is that of coherence, as defined by de Finetti (1937). An assessment of probabilities, interpreted as betting rates, is called *coherent* when a Dutch book—a combination of bets such that the person will always lose money—against a person who holds those assessments cannot be formed. It can be shown that a set of probability assessments is coherent if and only if it satisfies the usual axioms of probability calculus, including finite—but not countable—additivity and Bayes's rule (de Finetti 1937, Bernardo and Smith 1994).

The second property that we impose on a predictive distribution is open-mindedness, as defined in the previous section. A coherent predictive distribution that is also open-minded has been called *strictly coherent*, by Carnap (1952). An important feature of open-minded coherent predictive distributions is that they are completely determined by their marginal predictive distribution through Bayes's rule.

De Finetti (1937) also argued for the use of the third property that we impose: exchangeability. An infinite sequence of random quantities is called *exchangeable* when, for every finite subsequence, all permutations are considered equally probable. In other words, when a researcher considers a potentially infinite sequence of random quantities to be exchangeable, this means that he or she thinks that the order of the results in any given sample is irrelevant.

Running example. The predictive distribution that corresponds to the beta prior—see Equation (4)—is open-minded, coherent and exchangeable. Open-mindedness and coherence are obvious. For exchangeability, we look at the marginal predictive distribution (which, by coherence and open-mindedness, fully determines the predictive distribution)—see Equation (2). Clearly, this marginal predictive distribution does not depend on the sample order. $\Diamond$

Exchangeability plays a central role in relating coherent marginal predictive distributions to priors. The key to this relationship was demonstrated by de Finetti (1928, 1937) for (potentially) infinite sequences of binary random quantities considered to be exchangeable. His result can be extended to sequences of random quantities that assume values in any finite set, such as $\mathscr{X} = \{0, 1, \ldots, r-1\}$. In this generalized case, de Finetti's result states that when the marginal predictive distribution is coherent and exchangeable, there exist a random vector Θ that assumes values in the open standard $(r-1)$-simplex

$$\Delta_r := \left\{\theta = (\theta_1, \ldots, \theta_{r-1}) \in \mathbb{R}^{r-1} \colon \sum_{i=1}^{r-1} \theta_i < 1 \text{ and } (\forall i \in \{1, \ldots, r-1\})\ \theta_i > 0\right\}$$

with a unique distribution function $\Pi(\theta)$, such that

$$P_{\mathscr{X}}(\mathbf{S}_n = \mathbf{k}) = \int_{\Delta_r} \binom{n}{k_1, \ldots, k_{r-1}} \theta_1^{k_1} \theta_2^{k_2} \ldots (1 - \sum_{i=1}^{r-1} \theta_i)^{n - \Sigma_{i=1}^{r-1} k_i} d\Pi(\theta),$$

where $\mathbf{S}_n$ is the count vector that contains the number of times each category $j = 1, 2, \ldots, r-1$ was observed in $n \in \mathbb{N}$ trials and, therefore:

$$\mathbf{k} \in \left\{(k_1, \ldots, k_{r-1}) \in \mathbb{N}^{r-1} \colon \sum_{i=1}^{r-1} k_i \leq n\right\}.$$

Since, for any given sequence $(X_1 = e_1, \ldots, X_n = e_n) \in \mathscr{X}^n$ that has k_j instances of type j, for $j \in \{1, \ldots, r-1\}$, exchangeability and coherence imply that

$$P_{\mathscr{X}}(\mathbf{S}_n = \mathbf{k}) = \binom{n}{k_1, \ldots, k_{r-1}} P_{\mathscr{X}}(X_1 = e_1, \ldots, X_n = e_n),$$

we find that

$$P_{\mathscr{X}}(X_1 = e_1, \ldots, X_n = e_n) = \int_{\Delta_r} \theta_1^{k_1} \theta_2^{k_2} \ldots (1 - \sum_{i=1}^{r-1} \theta_i)^{n - \Sigma_{i=1}^{r-1} k_i} d\Pi(\theta);$$

Equation (1) corresponds to the special case in which $\mathscr{X}$ is binary, with $k = k_1$. Conversely, when your probability assessments can be written as a convex mixture of multinomial likelihoods, it is easy to show that they are exchangeable and coherent. This one-to-one correspondence is de Finetti's famous Representation Theorem.

Running example. When $\mathscr{X} = \{0,1\}$ is binary, $\mathbf{S}_n$ is a scalar that represents the counts of category "1" in n trials. The uniform distribution of $\mathbf{S}_n$ on the counts, $P_{\mathscr{X}}(\mathbf{S}_n = k) = 1/(n+1)$ for $k = 0,\dots,n$, is exchangeable and coherent and is implied by the uniform distribution on $(0,1)$—the beta prior with $\alpha = \beta = 1$—only; see (De Finetti 1928). Murray (1930) provides an alternative demonstration; see also (Geisser 1984). Similarly, the Polya distribution—see Equation (2)—can only be derived from the beta prior (Feller 1968, vol. II, p.229-230). $\diamondsuit$

By combining de Finetti's theorem with our assumption of open-mindedness, we obtain a one-to-one correspondence between predictive distributions and priors. The following theorem rephrases this result for future reference.

Theorem 1. *Every open-minded coherent exchangeable predictive distribution has a unique distribution function on the parameter space (a prior) from which it can be derived through the multinomial likelihood and Bayes's rule.*

By now, we have already justified the approach of combining a prior with a multinomial likelihood to obtain a predictive distribution. The next step is to reduce the set of possible priors by requiring the predictive distribution to satisfy additional principles, in addition to coherence, open-mindedness and exchangeability. This strategy was adopted by W. E. Johnson (1932) when he introduced the notion of predictive sufficiency, later called the *sufficientness postulate* by Good (1965). Together with exchangeability, Johnson used this postulate to justify the use of symmetric Dirichlet priors. We discuss Zabell's (2005, chap. 4) extension of Johnson's original approach, which moves beyond the symmetric case.

Definition 3. Let $P_{\mathscr{X}}$ be a coherent predictive distribution for random quantities assuming values in a finite category set $\mathscr{X}$. Then $P_{\mathscr{X}}$ is said to satisfy Johnson's sufficientness postulate if the immediate conditional predictive distribution is of the following form:

$$P_{\mathscr{X}}(X_{n+1} = j \mid X_1 = e_1, \ldots, X_n = e_n) = f_j(k_j, n), \tag{5}$$

where k_j is the number of times the outcome j was observed and $\sum_{j \in \mathscr{X}} k_j = n$.

In other words: the conditional probability that the next trial is of type j depends only on this type j, on the total number of observations, and on the number of times j was observed. The frequencies of other categories, or the specific trials in which category j appears, are not considered relevant.

Example 2. Let $\mathscr{X} = \{0, 1, \ldots, j, \ldots, r-1\}$ and consider a Dirichlet prior with hyperparameters $(\alpha_0, \ldots, \alpha_j, \ldots, \alpha_{r-1})$. The corresponding immediate conditional predictive distribution is then given by

$$P_{\mathscr{X}}(X_{n+1} = j \mid X_1 = e_1, \ldots, X_n = e_n) = \frac{k_j + \alpha_j}{n + \alpha} =: f_j(k_j, n),$$

which clearly satisfies Johnson's sufficientness postulate. ◊

If $|\mathscr{X}| > 2$, then any coherent predictive distribution that satisfies Equation (5) will be linear in k_j (Zabell 2005):

$$P_{\mathscr{X}}(X_{n+1} = j \mid X_1 = e_1, \ldots, X_n = e_n) = a_j(n) + b(n)k_j. \tag{6}$$

Furthermore, if an open-minded coherent predictive distribution is induced by a prior and the multinomial likelihood (if it is exchangeable), then Equation (6) implies that *the prior on the parameter space is a Dirichlet* or, when the random quantities are independent, a degenerate distribution (Zabell 2005).

The independent case is not particularly appealing in the context of predictive inference, in which the goal is to use past experience to provide useful information about future experiments of the same type. Indeed, when we assume that the experiments are i.i.d. with an 'unknown but fixed probability' p_j of resulting in category j, the predictive distribution will always provide p_j as its immediate conditional prediction for the probability of category j, thereby discarding the outcomes of previous trials. To avoid this behavior, we require that a predictive distribution allows a subject to learn from experience.

Definition 4. A predictive distribution is said to *allow a subject to learn from experience* if the observed data may provide relevant information about future experiments, i.e., there are $n \in \mathbb{N}$, $(e_1, \ldots, e_n) \in \mathscr{X}^n$ and $j \in \mathscr{X}$ such that

$$P_{\mathscr{X}}(X_{n+1} = j | X_1 = e_1, \ldots, X_n = e_n) \neq P_{\mathscr{X}}(X_{n+1} = j).$$

From a practical point of view, learning from experience is clearly a useful property, which is why we suggest imposing this constraint on predictive distributions. Because learning from experience implies that the experiments cannot be independent, the discussions in this section lead us to the following result.

Proposition 2. *If a subject accepts the open-mindedness condition (Definition 2) and has a coherent exchangeable predictive distribution $P_{\mathscr{X}}$, with $|\mathscr{X}| > 2$, that satisfies Johnson's sufficientness postulate (Definition 3) and allows learning from experience (Definition 4), the prior on the parameter space is a Dirichlet, from which the predictive distribution can be derived by combining the prior with the multinomial likelihood and applying Bayes's rule.*

This result is remarkable: by imposing simple principles on a predictive distribution, and without any other assumptions, we find that this predictive distribution is induced by a Dirichlet prior and the multinomial likelihood. Using a prior that is not a Dirichlet ensures that at least one of the imposed principles fails. However, the result only holds when $|\mathscr{X}| > 2$. In the binary case, linearity—Equation (6)—must be added as an assumption because it is no longer implied by Johnson's sufficientness postulate. This assumption weakens the result because linearity, unlike Johnson's sufficientness postulate, is not a philosophical principle.

4 From prediction systems to a family of priors

Because of de Finetti's representation theorem, solving the fundamental problem of practical statistics and its generalization to the non-binary case appears to have been reduced to selecting a prior. However, this is not entirely true. Before fixing a prior, the categories must be selected according to which the outcome of a single experiment will be classified. In other words, one has to partition what we will call the *possibility space* Ω, which is a set that contains all the possible outcomes of the experiment one could envision. Partitioning this possibility space corresponds to choosing a *sample space* $\mathscr{X}$, which is a set that contains the labels according to which the outcomes will be classified. In practice, the sample space will typically be finite, even when the possibility space is not.

Example 3. Suppose we are measuring the height of people, in meters. A reasonable possibility space for such an experiment would be $\Omega = (0,3]$,[3] which is infinite. However, because of the inherent limited accuracy of measurement methods (common measuring tools will show the height up to the closest 1 millimeter), the sample space $\mathscr{X}$ will be a finite partition of $(0,3]$. $\Diamond$

[3] Because the current Guinness Book World Record is 2.72 *m*.

Choosing a sample space—partitioning the possibility space—is often regarded as trivial. The most common approach is to base this choice either on the available data or on the particular inference problem at hand. However, in principle, there is no reason to do so, and one has to keep in mind that the choice of a partition can influence the resulting predictions. The following example illustrates this.

Example 4. Consider a disease from which both females (F) and males (M) can suffer and for which a treatment is available that may cure (C) or not cure ($\overline{C}$) the patient. Hence, $\mathscr{X} = \{CM, CF, \overline{C}M, \overline{C}F\}$ would be an obvious choice of sample space. The probabilities of the elements of $\mathscr{X}$ are designated by $\theta := (\theta_1, \ldots, \theta_4)$, which takes values in the standard open 3-simplex. Two physicians, working in different countries, each provide their prior opinion about θ in the form of a Dirichlet distribution. Their respective hyperparameters are $\alpha = (9,1,1,1)$ and $\beta = (1,1,1,9)$.

In a third country, a study tested the effect of the treatment on eight people affected by the disease. Amongst them, three men and one woman were cured and three men and one woman were not cured. We denote this dataset by $D = (3,1,3,1)$. Table 1 provides immediate conditional (one step ahead) predictions for the next patient. Three different priors are considered: $\mathrm{Dir}(\alpha)$, $\mathrm{Dir}(\beta)$ and a mixture that assigns equal weight to both.

	CM	CF	$\overline{C}M$	$\overline{C}F$	C	$\overline{C}$
$\mathrm{Dir}(\alpha)$	$3/5$	$1/10$	$1/5$	$1/10$	$7/10$	$3/10$
$\mathrm{Dir}(\beta)$	$1/5$	$1/10$	$1/5$	$1/2$	$3/10$	$7/10$
$1/2(\mathrm{Dir}(\alpha)+\mathrm{Dir}(\beta))$	$37/72$	$1/8$	$5/24$	$11/72$	$23/36$	$13/36$

Table 1: Immediate conditional predictions using the original data and priors

Consider now a situation in which no data are available about the sex of the previous patients, but only about whether or not they were cured. In that case, a natural way to proceed is to pool the original categories together and use $\mathscr{Y} := \{C, \overline{C}\}$ as the sample space. In this new sample space, the dataset is $D' = (4,4)$, and the priors are obtained by marginalizing the original priors: $\mathrm{Dir}(\alpha')$, with $\alpha' = (10,2)$, $\mathrm{Dir}(\beta')$, with $\beta' = (2,10)$, and the mixture that assigns equal weight to both. Table 2 provides the corresponding immediate conditional predictions for the next patient.

	C	$\overline{C}$
$\mathrm{Dir}(\alpha')$	$7/10$	$3/10$
$\mathrm{Dir}(\beta')$	$3/10$	$7/10$
$1/2(\mathrm{Dir}(\alpha') + \mathrm{Dir}(\beta'))$	$1/2$	$1/2$

Table 2: Immediate conditional predictions using the pooled data and priors

Suppose now that we are interested in the probability that the next patient will be cured, regardless of his or her sex, or equivalently, in the posterior expected value of $\theta_C := \theta_1 + \theta_2$. Intuitively, because this prediction problem is not concerned with the sex of the patient, one would expect that the result does not depend on whether we use the original sample space, data and priors or their pooled versions. However, as illustrated by the highlighted parts in Tables 1 and 2, such a dependency can occur when using mixtures of Dirichlets as a prior. No such behavior is observed for the individual Dirichlet priors in our example. $\lozenge$

In order to draw attention to this interaction between the sample space, the prior and the resulting inferences, we propose to look at inference from a somewhat distant perspective.

A predictive distribution for a particular partition of the possibility space is not a stand-alone inference tool but rather part of what we call a *predictive distribution system*.

Definition 5. A predictive distribution system Φ_Ω is a map from the set of all finite partitions of a possibility space Ω to the set of all predictive distributions. For every possible finite partition $\mathscr{X}$ of Ω—every finite sample space $\mathscr{X}$—a predictive distribution system Φ_Ω provides a corresponding predictive distribution $P_{\mathscr{X}}$.

Properties of predictive distributions can be imposed on predictive distribution systems, simply by imposing them element-wise. Using this convention, *in the remainder of this paper, we will consider every predictive distribution system to be coherent, open-minded and exchangeable*. Because of Theorem 1, this allows us to identify a predictive distribution system with a map from finite partitions to priors. For every finite partition $\mathscr{X}$ of the possibility space Ω, a predictive distribution system provides us with a corresponding prior.

Example 5. Consider a (naive) predictive distribution system Φ_Ω that assigns the uniform prior to every sample space. Let $\{A, B_1, B_2\}$ be a partition of Ω and define $B := B_1 \cup B_2$. Consider the sample spaces $\mathscr{X} := \{A, B\}$ and $\mathscr{Y} := \{A, B_1, B_2\}$, with the associated predictive distributions $P_{\mathscr{X}}$ and $P_{\mathscr{Y}}$. This predictive distribution system suffers from a problem: according to the predictive distribution $P_{\mathscr{X}}$, the marginal predictive probabilities of A and B are equal, whereas according to $P_{\mathscr{Y}}$, the marginal predictive probability of $B = B_1 \cup B_2$ is twice as large as that of A. $\lozenge$

As this example illustrates, some predictive distributions systems are clearly unreasonable because the marginal predictive probabilities of the different predictive distributions they consist of contradict each other. We will require that predictive distribution systems avoid this type of contradictions and will call any predictive distribution system that does so *marginally partition invariant*.

Definition 6. A predictive distribution system Φ_Ω is marginally partition invariant if for any finite partition $\mathscr{X}$ of the possibility space Ω, and any finite refinement or coarsening $\mathscr{Y}$ of $\mathscr{X}$, the marginal predictive distributions that correspond to $P_{\mathscr{X}}$ and $P_{\mathscr{Y}}$ do not contradict each other.

As illustrated by Example 4, requiring a predictive distribution system to be marginally partition invariant does not imply that all of the inferences are invariant to how the possibility space is partitioned. This requirement implies this invariance for the marginal predictive distributions—inferences drawn prior to observing any data—but the conditional predictive distributions may not exhibit such invariance. If the conditional predictive distributions are also invariant, then the predictive distribution system satisfies the *partition invariance principle*. We call such predictive distribution systems *partition invariant*.

Definition 7. A predictive distribution system Φ_Ω is partition invariant if for any finite partition $\mathscr{X}$ of the possibility space Ω, any finite refinement or coarsening $\mathscr{Y}$ of $\mathscr{X}$, and any—possibly empty—finite dataset that is sufficiently detailed to allow for the elements to be labeled according to both partitions $\mathscr{X}$ and $\mathscr{Y}$, the predictive distributions $P_{\mathscr{X}}$ and $P_{\mathscr{Y}}$ do not contradict each other: for any two labelings $(e_1,\ldots,e_n) \in \mathscr{X}^n$ and $(e_1^*,\ldots,e_n^*) \in \mathscr{Y}^n$ of the same—possibly empty—dataset and any future event E that can be expressed in terms of both $\mathscr{X}$ and $\mathscr{Y}$:

$$P_{\mathscr{X}}(E|X_1 = e_1,\ldots,X_n = e_n) = P_{\mathscr{Y}}(E|X_1 = e_1^*,\ldots,X_n = e_n^*). \tag{7}$$

In other words, predictive inferences drawn by a partition invariant predictive distribution system do not depend on how the possibility space is partitioned, thereby avoiding situations such as the one described in Example 4.

Predictive distribution systems that are derived from Dirichlet priors are prime examples of partition invariant predictive distribution systems.[4] The following example illustrates this for a ternary possibility space. Analogous examples can be constructed for other possibility spaces.

Example 6. Consider the possibility space $\Omega = \{0,1,2\}$ and the four partitions—sample spaces—that correspond to it: $\mathscr{X}_1 := \{\{0\},\{1\},\{2\}\}$, $\mathscr{X}_2 := \{\{0\},\{1,2\}\}$, $\mathscr{X}_3 := \{\{1\},\{0,2\}\}$ and $\mathscr{X}_4 := \{\{2\},\{0,1\}\}$. Consider α_0, α_1, $\alpha_2 > 0$ and let Φ_Ω be the predictive inference system that maps $\mathscr{X}_1$ to (the predictive distribution that corresponds to) $\mathrm{Dir}(\alpha_0,\alpha_1,\alpha_2)$, $\mathscr{X}_2$ to $\mathrm{Dir}(\alpha_0,\alpha_1+\alpha_2)$, $\mathscr{X}_3$ to $\mathrm{Dir}(\alpha_1,\alpha_0+\alpha_2)$ and $\mathscr{X}_4$ to $\mathrm{Dir}(\alpha_2,\alpha_0+\alpha_1)$. As we will show, this predictive inference system is partition invariant.

For reasons of symmetry, it suffices to prove Equation (7) for $\mathscr{X} = \mathscr{X}_1$ and $\mathscr{Y} = \mathscr{X}_2$. So consider two labelings $(e_1,\dots,e_n) \in \mathscr{X}^n$ and $(e_1^*,\dots,e_n^*) \in \mathscr{Y}^n$ of the same—possibly empty—dataset and any future event E that can be expressed in terms of both $\mathscr{X}$ and $\mathscr{Y}$. If $E = (X_{n+1} \in \{1,2\})$, it follows from Example 2 that

$$\begin{aligned} &P_{\mathscr{X}}(E|\{X_i = e_i\}_{i=1}^n) = \\ &\qquad = P_{\mathscr{X}}(X_{n+1} = 1|\{X_i = e_i\}_{i=1}^n) + P_{\mathscr{X}}(X_{n+1} = 2|\{X_i = e_i\}_{i=1}^n) \\ &\qquad = \frac{k_1+\alpha_1}{n+\alpha} + \frac{k_2+\alpha_2}{n+\alpha} = \frac{(k_1+k_2)+(\alpha_1+\alpha_2)}{n+\alpha} = P_{\mathscr{Y}}(E|\{X_i = e_i^*\}_{i=1}^n), \end{aligned}$$

with $\alpha := \alpha_0+\alpha_1+\alpha_2$. Similarly, for $E = (X_{n+1} = 0)$, we find that

$$P_{\mathscr{X}}(E|\{X_i = e_i\}_{i=1}^n) = \frac{k_0+\alpha_0}{n+\alpha} = P_{\mathscr{Y}}(E|\{X_i = e_i^*\}_{i=1}^n).$$

Hence, Equation (7) is true if E is an immediate event, in the sense that it only depends on the outcome of the next experiment. Since probabilities of more general

[4] In fact, we will prove further on that they are the only ones.

future events are completely characterized by conditional probabilities of immediate events, it follows that Φ_Ω is partition invariant. $\Diamond$

Principles that are similar to that of partition invariance have already been proposed by other authors. In an imprecise-probabilistic context—with lower and upper probabilities—Walley (1996) introduced the representation invariance principle. Similar to partition invariance, this principle requires that inferences should not depend on the sample space used. We prefer our terminology because it stresses that the sample space is a partition of a possibility space and that the invariance that is imposed is with respect to this partition. In a measure-theoretic setting, Böge and Möcks (1986) proposed the learn-merge invariance principle, which is also similar, and used this principle to characterize Dirichlet priors. The main differences with our approach are that they apply their principle to a single predictive distribution on a single sample space (instead of a predictive distribution system) and consider only mergers, no refinements. Consequently—and in contrast to our approach—their results do not apply to cases in which the sample space is binary: since a binary sample space cannot be merged (coarsened), their learn-merge invariance principle cannot be applied to such sample spaces.

We now investigate some consequences of imposing the partition invariance principle on a predictive distribution system. First of all: partition invariance implies marginal partition invariance, because the latter corresponds to the special case where no data are observed. Secondly, and more importantly: the partition invariance principle is closely related to Johnson's sufficientness postulate.

Proposition 3. *If a predictive distribution system Φ_Ω satisfies the partition invariance principle (Definition 7), then each of its predictive distributions $P_\mathscr{X}$ satisfies Johnson's sufficientness postulate (Definition 3).*

Proof. Consider any finite partition $\mathscr{X}$ of Ω and any $j \in \mathscr{X}$. If $\mathscr{X}$ is binary, let $\mathscr{Y} = \mathscr{X}$. Otherwise, let $\mathscr{Y}$ be a binary coarsening of $\mathscr{X}$ such that $j \in \mathscr{Y}$. Consider now any finite dataset $(e_1, \ldots, e_n) \in \mathscr{X}^n$ and let $(e_1^*, \ldots, e_n^*) \in \mathscr{Y}^n$ be the corresponding coarsened dataset. Since Φ_Ω is assumed to be partition invariant, we find

$$P_{\mathscr{X}}(X_{n+1} = j \mid X_1 = e_1, \ldots, X_n = e_n) = P_{\mathscr{Y}}(X_{n+1} = j \mid X_1 = e_1^*, \ldots, X_n = e_n^*).$$

Because $\mathscr{Y}$ is binary and $P_{\mathscr{Y}}$ is exchangeable, the final expression clearly depends only on j, k_j and n. Hence, $P_{\mathscr{X}}$ satisfies Johnson's sufficientness postulate; see Definition 3. □

By combining this result with Proposition 2, we obtain the following justification for the use of Dirichlet priors.

Theorem 4. *Consider any coherent, open-minded (Definition 2) and exchangeable predictive distribution system Φ_Ω, with $|\Omega| > 2$, that allows a subject to learn from experience (Definition 4) and satisfies the partition invariance principle (Definition 7). Then, for any finite—possibly binary—partition $\mathscr{X}$ of Ω, the corresponding predictive distribution $P_{\mathscr{X}}$ is derived from the multinomial likelihood and a prior on the parameter space that is Dirichlet.*

Proof. For $|\mathscr{X}| > 2$, this is an immediate consequence of Propositions 2 and 3. For $|\mathscr{X}| \leq 2$, consider any ternary refinement $\mathscr{Y}$ of $\mathscr{X}$; this is possible because $|\Omega| > 2$. Since $|\mathscr{Y}| > 2$, we can again use Propositions 2 and 3 to find that $P_{\mathscr{Y}}$ is derived from the multinomial likelihood and a prior on the parameter space that is Dirichlet. Therefore, the marginal part of $P_{\mathscr{X}}$ follows a Dirichlet-multinomial distribution. Additionally, because Φ_Ω is partition invariant and therefore marginally partition invariant, the marginal part of $P_{\mathscr{X}}$ is the restriction of the marginal part of $P_{\mathscr{Y}}$ to events that can be expressed in terms of the sample space $\mathscr{X}$. Therefore, because of (Johnson, Kotz and Balakrishnan 1997, Section 35.13.1), and (Basu and Pereira 1982), we know that the marginal part of $P_{\mathscr{X}}$ also follows a Dirichlet-multinomial distribution. The result then follows from the fact that the Dirichlet-multinomial distribution is derived from the multinomial likelihood and a prior on the parameter space that is Dirichlet. □

Example 4 provides a nice illustration of this theorem. The distribution system that uses mixtures of Dirichlets as priors displays a failure of partition invariance, whereas the others—which use single Dirichlet priors—do not. Theorem 4 guarantees that similar behavior will be observed for any prior that is not a Dirichlet and, in particular, for all mixtures of Dirichlets.

From a technical perspective, the importance of Theorem 4 is twofold. First, this theorem replaces Johnson's sufficientness postulate with the partition invariance principle. Although the former is shown to be implied by the latter, we think that the partition invariance principle is easier to justify. This is because it explicitly relates inferences with respect to different sample spaces; we consider Example 4 to be particularly convincing. Second, and unlike Proposition 2, this theorem does not require the rather annoying assumption that $|\mathscr{X}| > 2$. Instead, all we have to do is regard $\mathscr{X}$ as a partition of some possibility space Ω for which $|\Omega| > 2$. In other words: the sample space $\mathscr{X}$ can be binary as long as the possibility space Ω is not. Example 4 illustrates this: although the final probability of interest can—and often will—be stated in terms of a binary sample space, the possibility space contains (at least) four elements, thereby allowing the application of Theorem 4.

5 Concluding remarks

Our contribution shows that when a subject wishes to use a predictive distribution that (i) is coherent and open-minded, (ii) reflects a judgment of exchangeability, (iii) allows learning from experience and (iv) is part of a predictive distribution system that satisfies the partition invariance principle and whose possibility space—not to be confused with the sample space—contains more than two elements, then this predictive distribution must be derived from the multinomial likelihood and a prior on the parameter space that is Dirichlet. By carefully distinguishing between the sample space and the possibility space and by focusing on predictive distribution systems rather than predictive distributions, we were able to cover the binomial case as well.

For us, this is "*certainly a more principled approach to the problem of assigning a prior, in stark contrast to assuming the prior is Dirichlet purely for reasons of mathematical convenience*" (Zabell 2011). The partition invariance principle seems to be desirable in many applications of the multinomial Bayes's problem. However, as Johnson (1932) did, we would like to remark that it is the researcher's business to assess whether the principles here proposed are reasonable.

In cases in which our principles are deemed unreasonable, other principles could be considered, possibly leading to other types of priors. The logical approach has extensive literature discussing other principles of inductive inference. An important example is that of "analogy by similarity", which is concerned with cases in which the categories are ordered or related in some manner. In such cases, Johnson's postulate—and thus also partition invariance—is unsuitable. See (Skyrms 1993), (Festa 1996) and (Niiniluoto 1981) for more information about analogy by similarity. (Kuipers 1980) and (Niiniluoto 2011) provide surveys of the logical approach on inductive inference.

We envision two main directions of future research, both of which fall beyond the scope of the present study. First, other justifications for the use of Dirichlet priors have been suggested in the literature, some of which—for example (Costantini 1979)—were not mentioned yet. A detailed comparison with our approach should be conducted. Second, we believe that the partition invariance principle can be used to justify the use of Dirichlet processes as well.

Acknowledgements

Marcio Diniz was supported by FAPESP (São Paulo Research Foundation), under the project 2012/14764-0 and wishes to thank the SYSTeMs Research Group at Ghent University for its hospitality and support. Jasper De Bock is a PhD Fellow of the 'Research Foundation – Flanders' (FWO Vlaanderen) and wishes to acknowledge its financial support.

The authors also thank Sérgio Wechsler, Carlos A. B. Pereira and Luís Esteves for useful references.

References

Bayes, T. and Price, R. (1763), "An essay towards solving a problem in the doctrine of chances," *Philosophical Transactions of the Royal Society*, 53, 370–418.

Basu, D. and Pereira, C. A. B. (1982), "On the Bayesian analysis of categorical data: the problem of nonresponse," *Journal of Statistical Planning and Inference*, 6, 345–362.

Bedrick, E. J., Christensen, R. and Johnson, W. (1996), “A new perspective on priors for generalized linear models,” *Journal of the American Statistical Association*, 91, 1450–1460.

Berger, J. O., Bernardo, J. M. and Sun, D. (2009), “The formal definition of reference priors,” *The Annals of Statistics*, 37, 905–938.

Bernardo, J. M. (1979), “Reference Prior Distributions for Bayesian Inference,” *Journal of the Royal Statistical Society, Series B*, 41, 113–147.

Bernardo, J. M. and Smith, A. F. M. *Bayesian Theory*. Wiley, Chichester, 1994.

Böge, W. and Möcks, J. (1986), “Learn-merge invariance of priors: a characterization of the Dirichlet distributions and processes,” *Journal of Multivariate Analysis*, 18, 83–92.

Carnap, R. (1952) *The Continuum of Inductive Methods*, The University of Chicago Press.

Carnap, R. and Stegmüller, W. (1959), *Induktive Logik und Warscheinlichkeit*, Springer.

Costantini, D. (1979), “The relevance quotient,” *Erkenntnis*, 14, 149–157.

Diaconis, P. and Ylvisaker, D. (1979), “Conjugate priors for exponential families,” *The Annals of Statistics*, 7, 269–281.

Feller, W. (1968), *An Introduction to Probability Theory and its Applications*, Wiley.

Festa, R. (1996), “Analogy and exchangeability in predictive inferences,” *Erkenntnis*, 45, 229–252.

de Finetti, B. (1928), “Funzione caratteristica di un fenomeno aleatorio,” in *Atti del Congresso Internazionale dei Matematici*, Zanichelli, 179–190.

de Finetti, B. (1937), “La prévision: ses lois logiques, ses sources subjectives,” *Annales de l’Institut Henri Poincaré*, 7, 1–68. English translation in Kyburg and Smokler (1964).

Geisser, S. (1984), “On Prior Distributions for Binary Trials,” *The American Statistician*, 38, 244–247.

Good, I. J. (1965), *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, MIT Press.

Hume, D. (1888), *A Treatise of Human Nature*, edited by L. A. Selby-Bigge, Clarendon Press.

Jaynes, E. T. (2003), *Probability Theory*, Cambridge University Press.

Jeffreys, H. (1961), *Theory of Probability*, Oxford University Press.

Johnson, W. E. (1924), *Logic: Part III*, Cambridge University Press.

Johnson, W. E. (1932), “Probability: the deductive and inductive problems,” *Mind*, 41, 409–423.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997), *Discrete Multivariate Distributions*, Wiley.

Kass, R. E. and Wasserman, L. (1996), “The Selection of Prior Distributions by Formal Rules,” *Journal of the American Statistical Association*, 91, 1343–1370.

Keynes, J. M. (1921), *A Treatise on Probability*, Macmillan.

Kuipers, T. A. F. (1980), “A survey of inductive systems,” in *Studies in Inductive Logic and Probability* (Vol. II) , ed. R. C. Jeffrey, University of California Press, 183–192.

Kyburg Jr. H. E and Smokler, H. E. editors, (1964), *Studies in Subjective Probability*, Wiley.

Lindley, D. (1991), *Making Decisions*, Wiley.

Murray, F. H. (1930), “Note on a scholium of Bayes,” *Bulletin of the American Mathematical Society*, 36, 129-132.

Niiniluoto, I. (1981), “Analogy and inductive logic,” *Erkenntnis*, 21, 1-34.

Niiniluoto, I. (2011), “The development of the Hintikka program,” in *Handbook of the History of Logic* (Vol. 10), eds. D. M. Gabbay, S. Hartman, and J. Woods, Elsevier, 311–356.

O’Hagan, A., Buck, C., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D., Oakley, J. and Rakow, T. (2006), *Uncertain Judgements: Eliciting Experts’ Probabilities*, Wiley.

Pearson, K. (1920), “The Fundamental Problem of Practical Statistics,” *Biometrika*, 13, 1–16.

Raiffa, H. and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Clinton Press.

Robert, C. (2007), *The Bayesian Choice*, Springer.

Savage, L. J. (1954), *The Foundations of Statistics*. Dover.

Schervish, M. J. (1995), *Theory of Statistics*. Springer.

Seaman III, J. W., Seaman Jr., J. W. and Stamey, J. D. (2012), "Hidden dangers of specifying noninformative priors," *The American Statistician*, 66, 77–84.

Skyrms, B. (1993), "Analogy by similarity in hyper-Carnapian inductive logic," in *Philosophical Problems of the Internal and External Worlds. Essays in the Philosophy of Adolf Grünbaum*, University of Pittsburgh Press, 273–282.

Skyrms, B. (1996), "Carnapian inductive logic and Bayesian statistics," in *Statistics, Probability and Game Theory*, IMS Lecture Notes - Monograph Series (Vol. 30), 321–336.

Walley, P. (1996), "Inferences from multinomial data: learning about a bag of marbles," with discussion, *Journal of the Royal Statistical Society, Series B*, 58, 3–57.

Zabell, S. L. (2005), *Symmetry and Its Discontents: Essays on the History of Inductive Probability*, Cambridge Studies in Probability, Induction, and Decision Theory, Cambridge, UK, Cambridge University Press.

Zabell, S. L. (2011), "Carnap and the logic of inductive inference," in *Handbook of the History of Logic* (Vol. 10), eds. D. M. Gabbay, S. Hartman, and J. Woods, Elsevier, 265–309.