Characterising Dirichlet Priors

Marcio Diniz, Jasper De Bock and Arthur Van Camp

UFSCar & Ghent University

Introduction

Goal: justify the use of Dirichlet priors by imposing philosophical principles on predictive inference rules, namely:

1. Exchangeability;
2. Coherence;
3. Open-mindedness; and
4. Learning from experience.

We combine them with an additional requirement: the *partition invariance principle*, closely related to W. E. Johnson's sufficientness postulate.

From priors to predictions

Bayes (1763): given that we have observed k successes in the first n trials, what is the probability that the next trial will be a success (or a failure)? Question easily generalised to the multivariate case and to any (finite) number of future trials.

Definition 1 (Predictive inference rule)

A predictive inference rule, $R_{\mathcal{X}}$, gives, for any finite initial sequence of realised experiments $\{X_i = e_i\}_{i=1}^{n}$, a numerical prediction in $[0, 1]$ about any proposition on a finite number of future experiments.

From priors to predictions

Bayes' solution: (potentially) infinite sequence of binary random quantities assumed to be conditionally i.i.d. given θ, the probability of "success".

Bayesian paradigm: prior for θ, $\Pi(\theta)$ and, from it, derive the probability of any given sequence $(X_1 = e_1, \dots, X_n = e_n) \in \{0,1\}^n$, in which we observe k successes in n trials:

$$\int_0^1 \theta^k (1-\theta)^{n-k} d\Pi(\theta) = P(X_1 = e_1, \dots, X_n = e_n), \qquad (1)$$

From priors to predictions: open-mindedness

In order not to have problems when constructing predictive inference rules in this way, we impose the following condition.

Definition 2 (Open-mindedness)

If we consider that every finite sequence of realised experiments has probability greater than zero, that is:

$$P(X_1 = e_1, \ldots, X_n = e_n) > 0 \text{ for all } n < \infty \text{ and } (e_1, \ldots, e_n) \in \mathcal{X}^n, \tag{2}$$

then we are respecting what we call the open-mindedness condition.

Loosely speaking: any finite sequence of realised experiments is considered possible.

From priors to predictions: some history

- Laplace: extensive use of this approach (with uniform priors). His justification is nowadays called *principle of insufficient reason*, first proposed by Bernoulli (1713). Another modern name is *Bayes-Laplace postulate*.
  Corollary of such a prior: Laplace's rule of succession, the most criticized predictive probability on philosophical and mathematical grounds.
- Edgeworth (1884): one could use other priors than the uniform, every time experience points to them.
- G. F. Hardy (1889): probably the first to propose the beta prior for the binomial problem.

From priors to predictions: some history

In principle, one could use any possible distribution as a prior. The obvious and most important question then is: how to choose a prior?

- Mathematical convenience, leading to conjugate priors, disseminated specially by Raiffa and Schlaifer (1961) and justified (for the exponential family) on mathematical grounds by Diaconis and Ylvisaker (1979).
- MCMC era: the question received a lot of attention once more.

  Two main approaches: subjectivists (prior must reflect the personal opinion of experts, i.e. elicitation) "objectivists" (prior satisfies some structural criterion; reference priors—Bernardo (1979)— maximum entropy—Jaynes (2003)—invariance to reparametrisations—Jeffreys (1961)).

From predictions to priors

De Finetti (1937): parameters—on which the priors are defined—have no real operational meaning. Then, to **properly justify** the use of a specific prior, one should derive it by imposing properties on the predictive inference rule it produces. Did Bayes try this? Stigler (1982).

Another requirement: *coherence*. If a set of probability assessments is coherent, they satisfy the usual axioms of probability calculus, including finite, but not countable, additivity.

From predictions to priors: exchangeability

De Finetti (1937) introduced exchangeability: an infinite sequence of random quantities is considered *exchangeable* if for every finite subsequence, all permutations are equally probable.

Key result: for (potentially) infinite sequences of binary random quantities considered to be exchangeable, there is a random quantity Θ, that assumes values in $[0, 1]$, with unique distribution function $\Pi(\theta)$ such that:

$$P(S_n = k) = \int_0^1 \binom{n}{k} \theta^k (1-\theta)^{n-k} d\Pi(\theta) \tag{3}$$

where $S_n = X_1 + \ldots + X_n$, for every n and k, $n \in \mathbb{N}$ and $k \leq n$. This is the so-called de Finetti's Representation Theorem.

From predictions to priors

Theorem 3 (De Finetti representation theorem: dFRT)

If we respect the open-mindedness condition, a coherent and exchangeable predictive inference rule corresponds uniquely to a distribution function on the parameter space (prior) trough the multinomial likelihood and Bayes's theorem.

de Finetti (1928): the uniform distribution on the counts of n trials, $P(S_n = k) = 1/(n+1)$ for $k = 0, \ldots, n$, is implied only by the uniform distribution on $[0, 1]$;

Polya urn scheme: derived only by the Dirichlet prior.

From predictions to priors: sufficientness

Justify specific priors by imposing principles on the predictive inference rule: W. E. Johnson (1932) introduced the *sufficientness postulate*.

Definition 4 (Sufficientness postulate)

Let $R_{\mathcal{X}}$ be a coherent predictive inference rule for random quantities assuming values in a finite category set $\mathcal{X}$. Johnson's sufficientness postulate assumes that:

$$P(X_{n+1} = j | X_1 = e_1, \ldots, X_n = e_n) = f_j(n_j, n), \tag{4}$$

where n_j is the number of times the outcome j was observed.

From predictions to priors

If $|\mathcal{X}| > 2$, any coherent predictive inference rule that satisfies Eq. (4) is linear in n_j

$$P(X_{n+1} = j | X_1 = e_1, \ldots, X_n = e_n) = a_j(n) + b(n)n_j. \qquad (5)$$

Accepting open-mindedness + exchangeability: dFRT implies that the likelihood is multinomial and, thanks to the linearity of the predictive inference rule, *the prior on the parameter space should be Dirichlet* or, if the random quantities are considered independent, a degenerate distribution; Zabell (2005).

From predictions to priors: learning from experience

Independent case is not appealing when it is believed that past provides information about future experiments of the same kind.

Definition 5 (Learning from experience)

We say that a predictive inference rule allows a subject to learn from experience if observed data provides him with relevant information about future experiments.

From a practical point of view, it is a useful property, which is why we suggest to impose it on predictive inference rules.

From predictions to priors

Since learning from experience implies that the experiments cannot be (probabilistically) independent, the reasoning above leads to the following result.

Proposition 6 (Johnson & Zabell)

If a subject accepts the open-mindedness condition (Definition 2) and has a coherent exchangeable predictive inference rule $R_{\mathcal{X}}$*, with* $|\mathcal{X}| > 2$*, that satisfies Johnson's sufficientness postulate (Definition 4) and allows him to learn from experience (Definition 5), his prior on the parameter space is a Dirichlet.*

From inference systems to a family of priors

Due to dFRT: to solve Bayes's problem, just pick a prior.

But first, partition the *possibility space*, set that contains all possible outcomes of the experiment one could envision.

Partitioning this space corresponds to choose a *sample space* for the experiment or, choosing the labels according to which the outcomes will be classified.

But ... such a partition can have a major influence on the resulting inferences.

Example

Disease affecting females (F) and males (M).

Treatment may cure (C) or not cure ($\overline{C}$) it.

Obvious sample space: $\mathcal{X} = \{CM, CF, \overline{C}M, \overline{C}F\}$.

The probabilities of the elements of $\mathcal{X}$ are designated by $\theta = (\theta_1, \ldots, \theta_4)$, which takes values in the standard 3-simplex.

Two physicians provide their prior opinion about θ as a Dirichlet: $\alpha = (9, 1, 1, 1)$ and $\beta = (1, 1, 1, 9)$.

Example

Third researcher: study tested the effect of the treatment on eight people affected by the disease, being three men and one woman cured and three men and one woman not cured. We denote this data set by $D = (3, 1, 3, 1)$.

Table 1 provides immediate (one step ahead) predictions for the next patient.

Three different priors: Dir(α), Dir(β) and a mixture that assigns equal weight to both.

Example

	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 predictions using the original data and priors

Example

New situation: the object of interest is the probability of the next patient being cured or not.

New sample space: $\mathcal{Y} = \{C, \overline{C}\}$ (pool the old categories).
New data: $D' = (4, 4)$.
New priors: $\text{Dir}(\alpha')$, $\alpha' = (10, 2)$, and $\text{Dir}(\beta')$, $\beta' = (2, 10)$, and the mixture with equal weight to both.

	C	$\overline{C}$
$\text{Dir}(\alpha')$	$^7/_{10}$	$^3/_{10}$
$\text{Dir}(\beta')$	$^3/_{10}$	$^7/_{10}$
$^1/_2(\text{Dir}(\alpha') + \text{Dir}(\beta'))$	$^1/_2$	$^1/_2$

Table 2: Immediate predictions using the pooled data and priors

Inference systems

A predictive inference rule for some partition of the possibility space is not a stand-alone inference tool, but part of an *inference system*.

Definition 7 (Inference system)

An inference system Φ_Ω is a map from the set of all finite partitions of some possibility space Ω to a set of predictive inference rules. For every possible finite partition $\mathcal{X}$ of Ω or, equivalently, every finite sample space $\mathcal{X}$, it provides a corresponding predictive inference rule $R_{\mathcal{X}}$.

Inference systems

We consider only **coherent**, **open-minded** and **exchangeable** inference systems.

Due to dFRT: inference system $\leftrightarrow$ map from finite partitions to priors.

Therefore, for every finite partition $\mathcal{X}$ of the possibility space Ω, the inference system provides us a corresponding prior.

Inference systems: consistency

Definition 8 (Consistency)

An inference system Φ_Ω is consistent if for any finite partition $\mathcal{X}$ of the possibility space Ω, and any finite refinement or coarsening $\mathcal{Y}$ of $\mathcal{X}$, the prior beliefs about any proposition on a finite number of future experiments, as given by $R_\mathcal{X}$ and $R_\mathcal{Y}$, are related through marginalisation.

Consistent inference systems: due to uniqueness guaranteed by dFRT, priors that correspond to $\mathcal{X}$ and $\mathcal{Y}$ will be related through marginalisation.

Partition invariance

Requiring consistency from an inference system does not imply that inferences are invariant to how you partition the possibility space.

It only requires this invariance for inferences made **prior** to observing any data. Your posterior beliefs are still allowed to depend on the chosen partition.

If **posterior** beliefs are invariant as well, the inference system is said to satisfy the *partition invariance principle*.

Partition invariance

Definition 9 (Partition invariance)

An inference system Φ_Ω is partition invariant if for any finite partition $\mathcal{X}$ of the possibility space Ω, any finite refinement or coarsening $\mathcal{Y}$ of $\mathcal{X}$, and any finite data set that is detailed enough to allow for the data to be labelled according to both partition $\mathcal{X}$ and $\mathcal{Y}$, the resulting posterior beliefs about any proposition on any finite number of future experiments, as given by $R_\mathcal{X}$ and $R_\mathcal{Y}$, are related through marginalisation.

Partition invariance

Predictive inferences made by a partition invariant inference system do not depend on how one chooses to partition the possibility space, thereby avoiding situations such as the one described in the example.

Note that for a partition invariant inference system, due to the uniqueness guaranteed by dFRT, the posteriors that correspond to $\mathcal{X}$ and $\mathcal{Y}$ will also be related through marginalisation.

Other approaches

- Walley (1996): representation invariance principle. Like partition invariance, it requires that inferences should not depend on the sample space that is used.

- Böge and Möcks (1986): learn-merge invariance principle, also very similar, and used it to characterise Dirichlet priors. The main difference is that they apply the principle to a single inference rule on a single sample space and consider only mergers, and no refinements.

Partition invariance

Proposition 10

If an inference system Φ_Ω that satisfies the partition invariance principle (Definition 9), then each of its inference rules $R_\mathcal{X}$ satisfies Johnson's sufficientness postulate (Definition 4).

By combining this result with Proposition 6, we obtain the following intuitive characterisation of the Dirichlet distribution.

Main result

Theorem 11 (Characterisation of Dirchlet priors)

Consider any inference system Φ_Ω, with $|\Omega| > 2$, that allows a subject to learn from experience (Definition 5) and satisfies the partition invariance principle (Definition 9). Then for any finite partition $\mathcal{X}$ of Ω, including the binary ones, the corresponding inference rule $R_\mathcal{X}$ is derived from a prior on the parameter space that is Dirichlet.

Concluding remarks

Our contribution shows that if one whishes to use a predictive inference rule that:

1. reflects a judgement of exchangeability;
2. allows learning from experience;
3. is open-minded; and
4. is part of an inference system that satisfies the partition invariance principle

then the corresponding prior should be a Dirichlet.

Concluding remarks

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 (2009).

The partition invariance principle is desirable in several applications of the multinomial Bayes's problem.

However, we remark, as Johnson (1932), that it is the researcher's business to assess if the principles here proposed are reasonable.

References

- 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.
- de Finetti, B. (1937), "La prévision: ses lois logiques, ses sources subjectives," *Annales de l'Institut Henri Poincaré*, 7, 1–68.
- Johnson, W. E. (1932), "Probability: the deductive and inductive problems," *Mind*, 41, 409–423.
- Zabell, S. L. (2005), *Symmetry and Its Discontents: Essays on the History of Inductive Probability*, Cambridge University Press.
- Zabell, S. L. (2009), "Carnap and the logic of inductive inference," *Handbook of the History of Logic* (Vol. 10), Elsevier, pp. 265–309.