Why Beta Priors: Invariance-Based Explanation

Olga Kosheleva[1], Vladik Kreinovich[1], and Kittawit Autchariyapanitkul[2]

[1]University of Texas at El Paso
El Paso, Texas 79968, USA
olgak@utep.edu, vladik@utep.edu
[2]Faculty of Economics, Maejo University
Chiang Mai, Thailand, kittar3@hotmail.com

1. Formulation of the Problem

- In the Bayesian approach:
  - when we do not know the probability $p \in [0,1]$ of some event,
  - it is usually recommended to use a Beta prior distribution for p, with pdf

$$\rho(x) = c \cdot x^{\alpha-1} \cdot (1-x)^{\beta-1}.$$

- There have been numerous successful application of the use of the Beta distribution in the Bayesian approach.
- How can we explain this success?
- Why not use some other family of distributions located on the interval $[0,1]$?

2. Formulation of the Problem (cont-d)

- The need for such an explanation is especially important now, when the statistician community is:
  - replacing the traditional p-value techniques
  - with more reliable hypothesis testing methods.
- One such method is the Minimum Bayesian Factor (MBF) method.
- This method is based on Beta priors $\rho(x) = c \cdot x^a$ corresponding to $\beta = 1$.
- In this paper, we provide a natural explanation for these empirical successes.

3. Main Idea

- We want to find a natural prior distribution on the interval $[0, 1]$.
- This distribution should describe how frequently different probability values p appear.
- In determining this distribution, a natural idea to take into account is that:
  - in practice,
  - all probabilities are, in effect, conditional probabilities.
- We start with some class, and in this class, we find the corresponding frequencies.

4. Main Idea (cont-d)

- From this viewpoint:
  - we can start with the original probabilities and with their prior distribution,
  - or we can impose additional conditions and consider the resulting conditional probabilities.
- For example, in medical data processing, we may consider the probability that:
  - a patient with a certain disease
  - recovers after taking the corresponding medicine.
- We can consider this original probability.
- Alternatively, we can consider the conditional probability that a patient will recover.
- For example, the condition can be that the patient is at least 18 years old.

5. Main Idea (cont-d)

- We can impose many such conditions.
- We are looking for a universal prior, a prior that would describe all possible situations.
- So, it makes sense to consider priors for which:
  - after such a restriction,
  - we will get the exact same prior for the corresponding conditional probability.

6. Let Us Describe This Main Idea in Precise Terms

- In general, the conditional probability $P(A \mid B)$ has the form
$$P(A \mid B) = \frac{P(A \,\&\, B)}{P(B)}.$$
- Crudely speaking, this means that:
  - when we transition from the original probabilities to the new conditional ones,
  - we limit ourselves to the original probabilities which do not exceed some value $p_0 = P(B)$, and
  - we divide each original probability by p_0.
- In these terms, the above requirement takes the following form: for each $p_0 \in (0, 1)$,
  - if we limit ourselves to the interval $[0, p_0]$,
  - then the ratios p/p_0 should have the same distribution as the original one.

7. Resulting Definition

- *Let us assume that we have a probability distribution with probability density* $\rho(x)$ *on the interval* $[0, 1]$.
- *We say that this distribution is* invariant *if:*
  - *for each* $p_0 \in (0, 1)$,
  - *the ratio* x/p_0 *(restricted to the values* $x \le p_0$*) has the same distribution, i.e.:*

$$\rho(x/p_0 : x \le p_0) = \rho(x).$$

8. Main Result and Its Proof

- **Proposition.** *A probability distribution is invariant if and only if it has a form*

$$\rho(x) = c \cdot x^a \text{ for some } c \text{ and } a.$$

- **Proof.** The conditional probability density has the form

$$\rho(x/p_0 : x \le p_0) = C(p_0) \cdot \rho(x/p_0).$$

- Here, C is an appropriate constant depending on p_0.
- Thus, the invariance condition has the form

$$C(p_0) \cdot \rho(x/p_0) = \rho(x).$$

- By moving the term $C(p_0)$ to the right-hand side and denoting $\lambda \stackrel{\text{def}}{=} 1/p_0$ (so that $p_0 = 1/\lambda$), we get

$$\rho(\lambda \cdot x) = c(\lambda) \cdot \rho(x).$$

9. Proof (cont-d)

- We get $\rho(\lambda \cdot x) = c(\lambda) \cdot \rho(x)$, where we denoted

$$c(\lambda) \stackrel{\text{def}}{=} 1/C(1/\lambda).$$

- The probability density function is an integrable function – its integral is equal to 1.
- *Known:* all integrable solutions of the above functional equation has the form $\rho(x) = c \cdot x^a$ for some c, a.
- The proposition is thus proven.
- *Reminder:* these distributions – corr. to $\beta = 1$ – are used in the Bayesian approach to hypothesis testing.

10. How to Get a General Prior Distribution

- The above proposition describes the case when:
  - we have a single distribution
  - corresponding to a single piece of prior information.
- In practice, we may have many different pieces of information:
  - some of these pieces are about the probability p of the corresponding event E,
  - some may be about the probability $p' = 1 - p$ of the opposite event $\neg E$.
- According to the above Proposition, each piece of information about p can be described by the pdf $c_i \cdot x^{a_i}$.
- Similarly, each piece of information about $p' = 1 - p$ can be described by the probability density

$$c'_j \cdot x^{a'_j}.$$

11. General Case (cont-d)

- In terms of the original probability $p = 1 - p'$, this probability density has the form

$$c'_j \cdot (1-x)^{a'_j}.$$

- All these piece of information are independent.
- So, a reasonable idea is to multiply these probability density functions.
- After multiplication, we get a distribution of the type

$$c \cdot x^a \cdot (a-x)^{a'}, \text{ where } a = \sum_i a_i \text{ and } a' = \sum_j a'_j.$$

- This is exactly the Beta distribution – for $\alpha = a + 1$ and $\beta = a' + 1$.
- Thus, we have indeed justified the use of Beta priors.

12. Acknowledgments

- This work was supported by the Institute of Geodesy, Leibniz University of Hannover.
- It was also supported in part by the US National Science Foundation grants:
  - 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and
  - HRD-1242122 (Cyber-ShARE Center of Excellence).
- This paper was written when V. Kreinovich was visiting Leibniz University of Hannover.

Home Page

Title Page

◀◀ ▶▶

◀ ▶

Go Back

Full Screen

Close

Quit

13. Bibliography

- A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, Florida, 2013.
- A. Gelman and C. P. Robert, "The statistical crises in science", *American Scientist*, 2014, Vol. 102, No. 6, pp. 460–465.
- K. R. Kock, *Introduction to Bayesian Statistics*, Springer, 2007.
- H. T. Nguyen, "How to test without p-values", *Thailand Statistician*, 2019, Vol. 17, No. 2, pp. i-x.
- R. Page and E. Satake, "Beyond p-values and hypothesis testing: using the Minimum Bayes Factor to teach statistical inference in undergraduate introductory statistics courses", *Journal of Education and Learning*, 2017, Vol. 6, No. 4, pp. 254—266.

14. Bibliography (cont-d)

- R. L. Wasserstein and N. A. Lazar, "The ASA's statement on p-values: context, process, and purpose", *American Statistician*, 2016, Vol. 70, No. 2, pp. 129–133.