Prior probability

In Bayesian statistical inference, a **prior probability distribution**, often simply called the **prior**, of an uncertain quantity is the probability distribution that would express one's beliefs about this quantity before some evidence is taken into account. For example, the prior could be the probability distribution representing the relative proportions of voters who will vote for a particular politician in a future election. The unknown quantity may be a parameter of the model or a latent variable rather than an observable variable.

Bayes' theorem calculates the renormalized pointwise product of the prior and the likelihood function, to produce the *posterior probability distribution*, which is the conditional distribution of the uncertain quantity given the data.

Similarly, the **prior probability** of a random event or an uncertain proposition is the unconditional probability that is assigned before any relevant evidence is taken into account.

Priors can be created using a number of methods.[1](pp27–41) A prior can be determined from past information, such as previous experiments. A prior can be *elicited* from the purely subjective assessment of an experienced expert. An *uninformative prior* can be created to reflect a balance among outcomes when no information is available. Priors can also be chosen according to some principle, such as symmetry or maximizing entropy given constraints; examples are the Jeffreys prior or Bernardo's reference prior. When a family of *conjugate priors* exists, choosing a prior from that family simplifies calculation of the posterior distribution.

Parameters of prior distributions are a kind of *hyperparameter*. For example, if one uses a beta distribution to model the distribution of the parameter *p* of a Bernoulli distribution, then:

- p is a parameter of the underlying system (Bernoulli distribution), and
- α and β are parameters of the prior distribution (beta distribution); hence *hyper*parameters.

Hyperparameters themselves may have hyperprior distributions expressing beliefs about their values. A Bayesian model with more than one level of prior like this is called a hierarchical Bayes model.

Contents

Informative priors

An *informative prior* expresses specific, definite information about a variable. An example is a prior distribution for the temperature at noon tomorrow. A reasonable approach is to make the prior a normal distribution with expected value equal to today's noontime temperature, with variance equal to the day-to-day variance of atmospheric temperature, or a distribution of the temperature for that day of the year.

This example has a property in common with many priors, namely, that the posterior from one problem (today's temperature) becomes the prior for another problem (tomorrow's temperature); pre-existing evidence which has already been taken into account is part of the prior and, as more evidence accumulates, the posterior is determined largely by the evidence rather than any original assumption, provided that the original assumption admitted the possibility of what the evidence is suggesting. The terms "prior" and "posterior" are generally relative to a specific datum or observation.

Weakly informative priors

A *weakly informative prior* expresses partial information about a variable. An example is, when setting the prior distribution for the temperature at noon tomorrow in St. Louis, to use a normal distribution with mean 50 degrees Fahrenheit and standard deviation 40 degrees, which very loosely constrains the temperature to the range (10 degrees, 90 degrees) with a small chance of being below -30 degrees or above 130 degrees. The purpose of a weakly informative prior is for regularization, that is, to keep inferences in a reasonable range.

Uninformative priors

An **uninformative prior** or **diffuse prior** expresses vague or general information about a variable. The term "uninformative prior" is somewhat of a misnomer. Such a prior might also be called a *not very informative prior*, or an *objective prior*, i.e. one that's not subjectively elicited.

Uninformative priors can express "objective" information such as "the variable is positive" or "the variable is less than some limit". The simplest and oldest rule for determining a non-informative prior is the principle of indifference, which assigns equal probabilities to all possibilities. In parameter estimation problems, the use of an uninformative prior typically yields results which are not too different from conventional statistical analysis, as the likelihood function often yields more information than the uninformative prior.

Some attempts have been made at finding a priori probabilities, i.e. probability distributions in some sense logically required by the nature of one's state of uncertainty; these are a subject of philosophical controversy, with Bayesians being roughly divided into two schools: "objective Bayesians", who believe such priors exist in many useful situations, and "subjective Bayesians" who believe that in practice priors usually represent subjective judgements of opinion that cannot be rigorously justified (Williamson 2010). Perhaps the strongest arguments for objective Bayesianism were given by Edwin T. Jaynes, based mainly on the consequences of symmetries and on the principle of maximum entropy.

As an example of an a priori prior, due to Jaynes (2003), consider a situation in which one knows a ball has been hidden under one of three cups, A, B, or C, but no other information is available about its location. In this case a *uniform prior* of $p(A) = p(B) = p(C) = 1/3$ seems intuitively like the only reasonable choice. More formally, we can see that the problem remains the same if we swap around the labels ("A", "B" and "C") of the cups. It would therefore be odd to choose a prior for which a permutation of the labels would cause a change in our predictions about which cup the ball will be found under; the uniform prior is the only one which preserves this invariance. If one accepts this invariance principle then one can see that the uniform prior is the logically correct prior to represent this state of knowledge. This prior is "objective" in the sense of being the correct choice to represent a particular state of knowledge, but it is not objective in the sense of being an observer-independent feature of the world: in reality the ball exists under a particular cup, and it only makes sense to speak of probabilities in this situation if there is an observer with limited knowledge about the system.

As a more contentious example, Jaynes published an argument (Jaynes 1968) based on Lie groups that suggests that the prior representing complete uncertainty about a probability should be the Haldane prior $p^{-1}(1-p)^{-1}$. The example Jaynes gives is of finding a chemical in a lab and asking whether it will dissolve in water in repeated experiments. The Haldane prior[2] gives by far the most weight to $\boldsymbol{p=0}$ and $\boldsymbol{p=1}$, indicating that the sample will either dissolve every time or never dissolve, with equal probability. However, if one has observed samples of the chemical to dissolve in one experiment and not to dissolve in another experiment then this prior is updated to the uniform distribution on the interval [0, 1]. This is obtained by applying Bayes' theorem to the data set consisting of one observation of dissolving and one of not dissolving, using the above prior. The Haldane prior is an improper prior distribution (meaning that it has an infinite mass). Harold Jeffreys devised a systematic way for designing uninformative priors as e.g., Jeffreys prior $p^{-1/2}(1-p)^{-1/2}$ for the Bernoulli random variable.

Priors can be constructed which are proportional to the Haar measure if the parameter space X carries a natural group structure which leaves invariant our Bayesian state of knowledge (Jaynes, 1968). This can be seen as a generalisation of the invariance principle used to justify the uniform prior over the three cups in the example above. For example, in physics we might expect that an experiment will give the same results regardless of our choice of the origin of a coordinate system. This induces the group structure of the translation group on X, which determines the prior probability as a constant improper prior. Similarly, some measurements are naturally invariant to the choice of an arbitrary scale (e.g., whether centimeters or inches are used, the physical results should be equal). In such a case, the scale group is the natural group structure, and the corresponding prior on X is proportional to $1/x$. It sometimes matters whether we use the left-invariant or right-invariant Haar measure. For example, the left and right invariant Haar measures on the affine group are not equal. Berger (1985, p. 413) argues that the right-invariant Haar measure is the correct choice.

Another idea, championed by Edwin T. Jaynes, is to use the principle of maximum entropy (MAXENT). The motivation is that the Shannon entropy of a probability distribution measures the amount of information contained in the distribution. The larger the entropy, the less information is provided by the distribution. Thus, by maximizing the entropy over a suitable set of probability distributions on X, one finds the distribution that is least informative in the sense that it contains the least amount of information consistent with the constraints that define the set. For example, the maximum entropy prior on a discrete space, given only that the probability is normalized to 1, is the prior that assigns equal probability to each state. And in the continuous case, the maximum entropy prior given that the density is normalized with mean zero and unit variance is the standard normal distribution. The principle of *minimum cross-entropy* generalizes MAXENT to the case of "updating" an arbitrary prior distribution with suitable constraints in the maximum-entropy sense.

A related idea, reference priors, was introduced by José-Miguel Bernardo. Here, the idea is to maximize the expected Kullback–Leibler divergence of the posterior distribution relative to the prior. This maximizes the expected posterior information about X when the prior density is $p(x)$; thus, in some sense, $p(x)$ is the "least informative" prior about X. The reference prior is defined in the asymptotic limit, i.e., one considers the limit of the priors so obtained as the number of data points goes to infinity. In the present case, the KL divergence between the prior and posterior distributions is given by

$$\boldsymbol{KL = \int p(t) \int p(x \mid t) \log \frac{p(x \mid t)}{p(x)}\, dx\, dt.}$$

Here, $\boldsymbol{t}$ is a sufficient statistic for some parameter $\boldsymbol{x}$. The inner integral is the KL divergence between the posterior $\boldsymbol{p(x \mid t)}$ and prior $\boldsymbol{p(x)}$ distributions and the result is the weighted mean over all values of $\boldsymbol{t}$. Splitting the logarithm into two parts, reversing the order of integrals in the second part and noting that $\boldsymbol{\log[p(x)]}$ does not depend on $\boldsymbol{t}$ yields

$$\boldsymbol{KL = \int p(t) \int p(x \mid t) \log[p(x \mid t)]\, dx\, dt \; - \; \int \log[p(x)] \int p(t) p(x \mid t)\, dt\, dx.}$$

The inner integral in the second part is the integral over t of the joint density $p(x, t)$. This is the marginal distribution $p(x)$, so we have

$$KL = \int p(t) \int p(x \mid t) \log[p(x \mid t)] \, dx \, dt \; - \; \int p(x) \log[p(x)] \, dx.$$

Now we use the concept of entropy which, in the case of probability distributions, is the negative expected value of the logarithm of the probability mass or density function or $H(x) = -\int p(x) \log[p(x)] \, dx.$ Using this in the last equation yields

$$KL = -\int p(t) H(x \mid t) \, dt + \; H(x).$$

In words, KL is the negative expected value over t of the entropy of x conditional on t plus the marginal (i.e. unconditional) entropy of x. In the limiting case where the sample size tends to infinity, the Bernstein-von Mises theorem states that the distribution of x conditional on a given observed value of t is normal with a variance equal to the reciprocal of the Fisher information at the 'true' value of x. The entropy of a normal density function is equal to half the logarithm of $2\pi ev$ where v is the variance of the distribution. In this case therefore $H = \log \sqrt{2\pi e / [NI(x*)]}$ where N is the arbitrarily large sample size (to which Fisher information is proportional) and $x*$ is the 'true' value. Since this does not depend on t it can be taken out of the integral, and as this integral is over a probability space it equals one. Hence we can write the asymptotic form of KL as

$$KL = -\log[1 \sqrt{kI(x*)}] - \int p(x) \log[p(x)] \, dx.$$

where k is proportional to the (asymptotically large) sample size. We do not know the value of $x*$. Indeed, the very idea goes against the philosophy of Bayesian inference in which 'true' values of parameters are replaced by prior and posterior distributions. So we remove $x*$ by replacing it with x and taking the expected value of the normal entropy, which we obtain by multiplying by $p(x)$ and integrating over x. This allows us to combine the logarithms yielding

$$KL = -\int p(x) \log[p(x) / \sqrt{kI(x)}] \, dx.$$

This is a quasi-KL divergence ("quasi" in the sense that the square root of the Fisher information may be the kernel of an improper distribution). Due to the minus sign, we need to minimise this in order to maximise the KL divergence with which we started. The minimum value of the last equation occurs where the two distributions in the logarithm argument, improper or not, do not diverge. This in turn occurs when the prior distribution is proportional to the square root of the Fisher information of the likelihood function. Hence in the single parameter case, reference priors and Jeffreys priors are identical, even though Jeffreys has a very different rationale.

Reference priors are often the objective prior of choice in multivariate problems, since other rules (e.g., Jeffreys' rule) may result in priors with problematic behavior.

Objective prior distributions may also be derived from other principles, such as information or coding theory (see e.g. minimum description length) or frequentist statistics (see frequentist matching). Such methods are used in Solomonoff's theory of inductive inference. Constructing objective priors have been recently introduced in bioinformatics, and specially inference in cancer systems biology, where sample size is limited

and a vast amount of **prior knowledge** is available. In these methods, either an information theory based criterion, such as KL divergence or log-likelihood function for binary supervised learning problems[3] and mixture model problems.[4]

Philosophical problems associated with uninformative priors are associated with the choice of an appropriate metric, or measurement scale. Suppose we want a prior for the running speed of a runner who is unknown to us. We could specify, say, a normal distribution as the prior for his speed, but alternatively we could specify a normal prior for the time he takes to complete 100 metres, which is proportional to the reciprocal of the first prior. These are very different priors, but it is not clear which is to be preferred. Jaynes' often-overlooked method of transformation groups can answer this question in some situations.[5]

Similarly, if asked to estimate an unknown proportion between 0 and 1, we might say that all proportions are equally likely, and use a uniform prior. Alternatively, we might say that all orders of magnitude for the proportion are equally likely, the **logarithmic prior**, which is the uniform prior on the logarithm of proportion. The Jeffreys prior attempts to solve this problem by computing a prior which expresses the same belief no matter which metric is used. The Jeffreys prior for an unknown proportion p is $p^{-1/2}(1-p)^{-1/2}$, which differs from Jaynes' recommendation.

Priors based on notions of algorithmic probability are used in inductive inference as a basis for induction in very general settings.

Practical problems associated with uninformative priors include the requirement that the posterior distribution be proper. The usual uninformative priors on continuous, unbounded variables are improper. This need not be a problem if the posterior distribution is proper. Another issue of importance is that if an uninformative prior is to be used *routinely*, i.e., with many different data sets, it should have good frequentist properties. Normally a Bayesian would not be concerned with such issues, but it can be important in this situation. For example, one would want any decision rule based on the posterior distribution to be admissible under the adopted loss function. Unfortunately, admissibility is often difficult to check, although some results are known (e.g., Berger and Strawderman 1996). The issue is particularly acute with hierarchical Bayes models; the usual priors (e.g., Jeffreys' prior) may give badly inadmissible decision rules if employed at the higher levels of the hierarchy.

Improper priors

Let events $A_1, A_2, \ldots, A_n$ be mutually exclusive and exhaustive. If Bayes' theorem is written as

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_j P(B \mid A_j)P(A_j)},$$

then it is clear that the same result would be obtained if all the prior probabilities $P(A_i)$ and $P(A_j)$ were multiplied by a given constant; the same would be true for a continuous random variable. If the summation in the denominator converges, the posterior probabilities will still sum (or integrate) to 1 even if the prior values do not, and so the priors may only need to be specified in the correct proportion. Taking this idea further, in many cases the sum or integral of the prior values may not even need to be finite to get sensible answers for the posterior probabilities. When this is the case, the prior is called an **improper prior**. However, the posterior distribution need not be a proper distribution if the prior is improper. This is clear from the case where event B is independent of all of the A_j.

Statisticians sometimes[6] use improper priors as uninformative priors. For example, if they need a prior distribution for the mean and variance of a random variable, they may assume $p(m, v) \sim 1/v$ (for $v > 0$) which would suggest that any value for the mean is "equally likely" and that a value for the positive variance becomes "less likely" in inverse proportion to its value. Many authors (Lindley, 1973; De Groot, 1937; Kass

and Wasserman, 1996) warn against the danger of over-interpreting those priors since they are not probability densities. The only relevance they have is found in the corresponding posterior, as long as it is well-defined for all observations. (The Haldane prior is a typical counterexample.)

By contrast, likelihood functions do not need to be integrated, and a likelihood function that is uniformly 1 corresponds to the absence of data (all models are equally likely, given no data): Bayes' rule multiplies a prior by the likelihood, and an empty product is just the constant likelihood 1. However, without starting with a prior probability distribution, one does not end up getting a posterior probability distribution, and thus cannot integrate or compute expected values or loss. See Likelihood function § Non-integrability for details.

Examples

Examples of improper priors include:

- The uniform distribution on an infinite interval (i.e., a half-line or the entire real line).
- Beta(0,0), the beta distribution for $\alpha=0$, $\beta=0$ (uniform distribution on log-odds scale).
- The logarithmic prior on the positive reals (uniform distribution on log scale).

Note that these functions, interpreted as uniform distributions, can also be interpreted as the likelihood function in the absence of data, but are not proper priors.

Notes

1. Carlin, Bradley P.; Louis, Thomas A. (2008). *Bayesian Methods for Data Analysis* (Third ed.). CRC Press. ISBN 9781584886983.
2. This prior was proposed by J.B.S. Haldane in "A note on inverse probability", Mathematical Proceedings of the Cambridge Philosophical Society 28, 55–61, 1932, doi:10.1017/S0305004100010495 (https://doi.org/10.1017%2FS0305004100010495). See also J. Haldane, "The precision of observed values of small frequencies", Biometrika, 35:297–300, 1948, doi:10.2307/2332350 (https://doi.org/10.2307%2F2332350), JSTOR 2332350 (https://www.jstor.org/stable/2332350).
3. Esfahani, M. S.; Dougherty, E. R. (2014). "Incorporation of Biological Pathway Knowledge in the Construction of Priors for Optimal Bayesian Classification - IEEE Journals & Magazine". *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. **11** (1): 202–18. doi:10.1109/TCBB.2013.143 (https://doi.org/10.1109%2FTCBB.2013.143). PMID 26355519 (https://pubmed.ncbi.nlm.nih.gov/26355519).
4. Boluki, Shahin; Esfahani, Mohammad Shahrokh; Qian, Xiaoning; Dougherty, Edward R (December 2017). "Incorporating biological prior knowledge for Bayesian learning via maximal knowledge-driven information priors" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5751802). *BMC Bioinformatics*. **18** (S14): 552. doi:10.1186/s12859-017-1893-4 (https://doi.org/10.1186%2Fs12859-017-1893-4). ISSN 1471-2105 (https://www.worldcat.org/issn/1471-2105). PMC 5751802 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5751802). PMID 29297278 (https://pubmed.ncbi.nlm.nih.gov/29297278).
5. Jaynes (1968), pp. 17, see also Jaynes (2003), chapter 12. Note that chapter 12 is not available in the online preprint but can be previewed via Google Books.
6. Christensen, Ronald; Johnson, Wesley; Branscum, Adam; Hanson, Timothy E. (2010). *Bayesian Ideas and Data Analysis : An Introduction for Scientists and Statisticians*. Hoboken: CRC Press. p. 69. ISBN 9781439894798.

References

- Rubin, Donald B.; Gelman, Andrew; John B. Carlin; Stern, Hal (2003). *Bayesian Data Analysis* (2nd ed.). Boca Raton: Chapman & Hall/CRC. ISBN 978-1-58488-388-3. MR 2027492 (https://www.ams.org/mathscinet-getitem?mr=2027492).
- Berger, James O. (1985). *Statistical decision theory and Bayesian analysis*. Berlin: Springer-Verlag. ISBN 978-0-387-96098-2. MR 0804611 (https://www.ams.org/mathscinet-getitem?mr=0804611).
- Berger, James O.; Strawderman, William E. (1996). "Choice of hierarchical priors: admissibility in estimation of normal means" (https://doi.org/10.1214%2Faos%2F1032526950). *Annals of Statistics*. **24** (3): 931–951. doi:10.1214/aos/1032526950 (https://doi.org/10.1214%2Faos%2F1032526950). MR 1401831 (https://www.ams.org/mathscinet-getitem?mr=1401831). Zbl 0865.62004 (https://zbmath.org/?format=complete&q=an:0865.62004).
- Bernardo, Jose M. (1979). "Reference Posterior Distributions for Bayesian Inference". *Journal of the Royal Statistical Society, Series B*. **41** (2): 113–147. JSTOR 2985028 (https://www.jstor.org/stable/2985028). MR 0547240 (https://www.ams.org/mathscinet-getitem?mr=0547240).
- James O. Berger; José M. Bernardo; Dongchu Sun (2009). "The formal definition of reference priors". *Annals of Statistics*. **37** (2): 905–938. arXiv:0904.0156 (https://arxiv.org/abs/0904.0156). Bibcode:2009arXiv0904.0156B (https://ui.adsabs.harvard.edu/abs/2009arXiv0904.0156B). doi:10.1214/07-AOS587 (https://doi.org/10.1214%2F07-AOS587).
- Jaynes, Edwin T. (Sep 1968). "Prior Probabilities" (http://bayes.wustl.edu/etj/articles/prior.pdf) (PDF). *IEEE Transactions on Systems Science and Cybernetics*. **4** (3): 227–241. doi:10.1109/TSSC.1968.300117 (https://doi.org/10.1109%2FTSSC.1968.300117). Retrieved 2009-03-27.
  - Reprinted in Rosenkrantz, Roger D. (1989). *E. T. Jaynes: papers on probability, statistics, and statistical physics*. Boston: Kluwer Academic Publishers. pp. 116–130. ISBN 978-90-277-1448-0.
- Jaynes, Edwin T. (2003). *Probability Theory: The Logic of Science* (http://www-biba.inrialpes.fr/Jaynes/prob.html). Cambridge University Press. ISBN 978-0-521-59271-0.
- Williamson, Jon (2010). "review of Bruno di Finetti. Philosophical Lectures on Probability" (https://web.archive.org/web/20110609175653/http://www.kent.ac.uk/secl/philosophy/jw/2009/deFinetti.pdf) (PDF). *Philosophia Mathematica*. **18** (1): 130–135. doi:10.1093/philmat/nkp019 (https://doi.org/10.1093%2Fphilmat%2Fnkp019). Archived from the original (http://www.kent.ac.uk/secl/philosophy/jw/2009/deFinetti.pdf) (PDF) on 2011-06-09. Retrieved 2010-07-02.
