

$$\frac{p(w) \pi(w)}{Z}, \quad (6.13)$$

theorem can help the reader better understand the expressions is returned in

be non-negative but are otherwise

$$\begin{aligned} &< \infty \\ &< \infty \\ &< \infty. \end{aligned} \quad (6.14)$$

densities locally may have values $p(w_*) = 0$ or $\pi(w_*) = 0$ is possible, and $P(w_*) = 0$. We will encounter an probability density is set to zero in

may range over $(-\infty, \infty)$. The p are restricted by the limitations always finite.

Bayes' theorem

s point. In classical statistics, the while the model parameters, such determines its value. Conceptually, parameter. However, in Bayesian e data are fixed, while the models ons of their parameters. hat a probability distribution is classical statistics, which implies experiment and convergence to inity. In Bayesian statistics, there ative experiments. Furthermore, Bayesian theory, only probability

Chapter 7

Probability densities

[P]rogress has to be made by invoking suitable assumptions; the simplest are best to begin with ... (Devinder Sivia)

7.1 Introduction

You may have already presumed there is no straightforward recipe for assigning a prior probability. And this notion is correct. Prior probabilities are *subjective*, but they are *not* arbitrary. A prior probability distribution is chosen after an informed judgment of the relevance and usefulness of all available information.

This chapter provides a few examples of probability density distributions that are often used as prior probabilities, shown in Figure 7.1.

7.2 Gaussian distribution

The best-known probability density is the univariate Gaussian distribution

$$p(w) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(w-\mu)^2}{2\sigma^2}\right). \quad (7.1)$$

In this expression, $p(w)$ represents the probability density at w , separated from its *mean* μ by $(w-\mu)$. The Gaussian distribution has infinite *support* $w \in (-\infty, \infty)$. The attenuation of the density is parameterized by the *standard deviation* σ , with $\sigma > 0$. All probability densities are non-negative, hence $p(w) \in [0, \infty)$. Notice the quadratic dependency of $(w-\mu)$ in the exponential. The mean of the Gaussian distribution coincides with

its maximum or *mode*. The familiar “bell-shaped” distribution is shown in Figure 7.1, Panel (a).

One may be surprised to learn that the Gaussian distribution is a very “compact” distribution, still having infinitely wide, very “thin tails.” Most probability mass is concentrated within a few standard deviations around the mean. Therefore, a Gaussian prior may be an appropriate choice when we want to confine a parameter w to a compact region.

7.3 Exponential distributions

The Laplace distribution is the two-sided exponential probability distribution

$$p(w) = \frac{1}{2c} \exp\left(-\frac{|w - \mu|}{c}\right). \quad (7.2)$$

Its mean is at μ , and the *rate parameter* is c , which is strictly positive. This distribution also has infinite support $w \in (-\infty, \infty)$. The tails of this distribution depend linearly on the separation $|w - \mu|$ and are much broader than those of the Gaussian distribution. However, unlike the Gaussian distribution, the Laplace distribution has a sharp cusp at $w = \mu$ as seen in Figure 7.1, Panel (b).

The one-sided exponential distribution is defined by

$$p(w) = \begin{cases} \frac{1}{\mu} \exp\left(-\frac{w}{\mu}\right) & w \geq 0 \\ 0 & w < 0, \end{cases} \quad (7.3)$$

where μ is the *mean* of the distribution, with $\mu > 0$. Also, the one-sided exponential distribution has a sharp cusp.

7.4 Cauchy distribution

The Cauchy distribution is

$$p(w) = \frac{s}{\pi(s^2 + w^2)}. \quad (7.4)$$

The s is a positive scale parameter, $s > 0$, and $w \in (-\infty, \infty)$. The Cauchy distribution is known to have “fat tails,” which attenuate slowly with w .

The one-sided half-Cauchy distribution is defined for positive values of w only

7.5 Uniform distribution

$$p(w) = \begin{cases} \frac{1}{u - l} & l \leq w \leq u \\ 0 & \text{otherwise} \end{cases}$$

The Cauchy distribution can be a useful distribution when dealing with ill-defined sources of data. The Cauchy distribution complements the Gaussian distribution to confine the parameter values. It is also called the Lorentz distribution.

Classical statisticians want to have a well-defined limit; it keeps “jumping” around. When sampling from a distribution, compute the sample mean $\bar{v}$ (5.6) and then, by adding ever more samples, the sample mean converges to a well-defined limit; it keeps “jumping” around.

This very counterintuitive result is discussed by Nolan ([8], p. 26). In Bayesian data analysis, the Cauchy distribution is pathological about the Cauchy distribution.

From the distributions presented in this chapter, the Gaussian and the Cauchy distribution are the same family of probability distributions. The Cauchy distribution is named Lévy distributions and the Gaussian distribution is named stable distributions are distributions that remain the same when adding or averaging variables.

7.5 Uniform distribution

The uniform probability distribution

$$p(w) = \begin{cases} \frac{1}{u - l} & l \leq w \leq u \\ 0 & \text{otherwise} \end{cases}$$

The uniform distribution has strict support in the line segment $w \in [l, u]$. In two dimensions, a volume, and in three dimensions, a volume, and so on.

A uniform prior is often assigned to a parameter that represents the possible values in an interval. The uniform distribution represents a phase angle that occupies the entire range of possible values.

"bell-shaped" distribution is shown

that the Gaussian distribution is a ring infinitely wide, very "thin tails" and within a few standard deviations Gaussian prior may be an appropriate parameter w to a compact region.

tions

ided exponential probability distribution

$$\left(-\frac{|w - \mu|}{c}\right). \quad (7.2)$$

eter is c , which is strictly positive support $w \in (-\infty, \infty)$. The tails of the separation $|w - \mu|$ and are much distribution. However, unlike the distribution has a sharp cusp at $w = \mu$.

tion is defined by

$$\begin{aligned} -\frac{w}{\mu} & \quad w \geq 0 \\ & \quad w < 0, \end{aligned} \quad (7.3)$$

n, with $\mu > 0$. Also, the one-sided distribution.

$$\frac{s}{s^2 + w^2}. \quad (7.4)$$

l, and $w \in (-\infty, \infty)$. The Cauchy distribution which attenuate slowly with w . The distribution is defined for positive values

$$p(w) = \begin{cases} \frac{2s}{\pi(s^2 + w^2)} & w \geq 0 \\ 0 & w < 0 \end{cases} \quad (7.5)$$

The Cauchy distribution can be adopted as a prior probability distribution when dealing with ill-defined situations. Its wide tails discriminate reasonably against outliers without prohibiting them. In this respect, the Cauchy distribution complements the Gaussian distribution, which tends to confine the parameter values. In physics, the Cauchy distribution is also called the Lorentz distribution.

Classical statisticians want to have nothing to do with the Cauchy distribution. When sampling from a Cauchy distribution, one can always compute the sample mean $\bar{v}$ (5.6) and the sample variance s_v^2 (5.7). However, by adding ever more samples, their mean does not converge to a well-defined limit; it keeps "jumping around." The same is true for the variance. This very counterintuitive result is nicely illustrated in the book by Nolan ([8], p. 26). In Bayesian data analysis, however, there is nothing pathological about the Cauchy distribution.

From the distributions presented so far, one may surmise that the Gaussian and the Cauchy distributions are perhaps the members of the same family of probability distributions. And indeed so, this family goes by the name Lévy distributions and the more general *stable distributions*. Stable distributions are distributions whose form (shape) remains the same when adding or averaging variables ([8]).

7.5 Uniform distribution

The uniform probability distribution is

$$p(w) = \begin{cases} \frac{1}{u - l} & w \in [l, u] \\ 0 & \text{otherwise.} \end{cases} \quad (7.6)$$

The uniform distribution has strict lower and upper bounds; its support is the line segment $w \in [l, u]$. In two dimensions, the support is an area; in three dimensions, a volume, and so on.

A uniform prior is often assigned when one cannot distinguish between the possible values in an interval. For example, if the parameter w represents a phase angle that occupies a full circle, $0 \leq w \leq 2\pi$.

The uniform distribution is often applied to situations where $[l, u]$ represents a segment on the real line, with no preferred value for w within this segment. However, on the one hand, this assumes that we are ignorant of any value of w , while, on the other hand, we seem to know precisely its boundary values. This is a contradiction in our state of knowledge. Nevertheless, uniform distributions on segments of the real line are used by many authors and also in this tutorial (Chapter 14). One should be aware that a uniform prior is a simplifying but sometimes helpful artifice.

The uniform prior is sometimes referred to as a *location prior*.

7.6 Jeffreys distribution

The British geophysicist Sir Harold Jeffreys (1891 - 1989) defined a suitable distribution for a *scale factor*, which now bears his name ([5], pg. 188; [4], pg. 181). A scale factor is a strictly positive and finite variable, for example, the standard deviation σ of a Gaussian distribution. The Jeffreys distribution is

$$p(w) = \begin{cases} \frac{c}{w} & w \in [l, u] \\ 0 & \text{otherwise.} \end{cases} \quad (7.7)$$

The normalization requires that $w > 0$ and that it lies within a bounded interval. The normalization constant is

$$\begin{aligned} c &= \left(\int_l^u \frac{1}{w} dw \right)^{-1} \\ &= \frac{1}{\ln(u/l)}. \end{aligned} \quad (7.8)$$

Another way of interpreting Jeffreys distribution is that it is a uniform distribution or the *logarithm* of w over the support $[l, u]$, or

$$p(\ln w) = \text{constant.} \quad (7.9)$$

A Jeffreys distribution is suitable for any parameter w representing a scale factor.

Again one has to impose strict limits on the interval. And the same discussion as in Section 7.5 applies to Jeffreys distribution. Nevertheless, *Jeffreys prior* is widely applied in many applications of Bayesian data

7.7 Conjugate priors

analysis. We will encounter applications in Chapter 14.

7.7 Conjugate priors

Prior distributions which belong to the same family as the likelihood are called *conjugate priors*. It is always multiplied by the likelihood from the same mathematical family, thereby often simplifying the calculations.

In this tutorial we will extend the Gaussian distribution. A Gaussian likelihood yields a Gaussian posterior that we can do all calculations analytically.

Conjugate priors, however, have just pragmatic choices that can be found in a substantial list of conjugate probability distributions (Wikipedia ([12])).

7.8 In practice

Experience in data analysis and the choice of the prior. In contrast, in the choice of your prior distribution often significant. In that case, you have made a conscious choice. And your time may be best spent on the choice of the prior.

We will see that the prior plays a role in the analysis, namely, stabilizing the estimates of the nuisance variables (Chapter 11), and in the choice of the prior.

The prior probability distribution is most used in practice. For more information see the books by Sivian and Skilling ([9], pg. 372), and Von der Linden, D.

applied to situations where $[l, u]$ with no preferred value for w within , this assumes that we are ignorant hand, we seem to know precisely fiction in our state of knowledge segments of the real line are used rial (Chapter 14). One should be ing but sometimes helpful artifice erred to as a *location prior*.

ys (1891 - 1989) defined a suitable now bears his name ([5], pg. 188; y positive and finite variable, for Gaussian distribution. The Jeffreys

$$v \in [l, u] \tag{7.7}$$

therwise.

and that it lies within a bounded

$$w)^{-1} \tag{7.8}$$

tribution is that it is a uniform the support $[l, u]$, or

$$\text{stant.} \tag{7.9}$$

parameter w representing a scale

s on the interval. And the same ffreys distribution. Nevertheless, y applications of Bayesian data

analysis. We will encounter applications of Jeffreys prior in Chapters 13 and 14.

7.7 Conjugate priors

Prior distributions which belong to the same family as the likelihood function are called *conjugate priors*. In Bayes' theorem, the prior probability is always multiplied by the likelihood. When both distributions come from the same mathematical family, the posterior also belongs to this family, thereby often simplifying the calculations.

In this tutorial we will extensively use the conjugate property of the Gaussian distribution. A Gaussian prior multiplied by a Gaussian likelihood yields a Gaussian posterior distribution, with the advantage that we can do all calculations analytically.

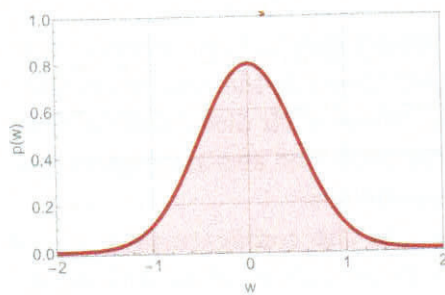
Conjugate priors, however, have *no* fundamental meaning. They are just pragmatic choices that can save one a lot of time and effort. A substantial list of conjugate probability distributions can be found on Wikipedia ([12]).

7.8 In practice

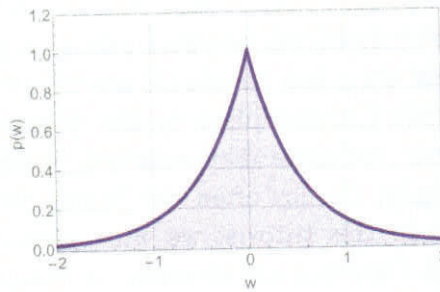
Experience in data analysis and domain understanding leads to a sound choice of the prior. In contrast, in an ill-defined problem, small changes to your prior distribution often significantly affect your numerical results. In that case, you have made a conceptual or, more likely, a programming error. And your time may be best spent re-stating your problem.

We will see that the prior plays three different roles in Bayesian data analysis, namely, stabilizing the overfitting of models and eliminating nuisance variables (Chapter 11), and model selection (Chapter 12).

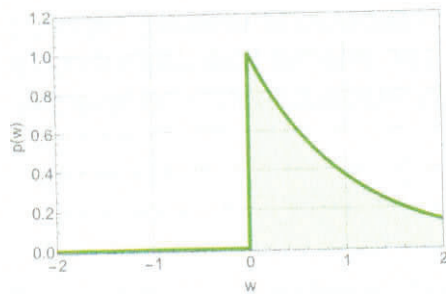
The prior probability distributions described in this chapter are the most used in practice. For more examples and discussion, consult the books by Sivia and Skilling ([9], pg. 117), Gregory ([3], pg. 184), Jaynes ([4], pg. 372), and Von der Linden, Dose, and Von Toussaint ([11], pg. 165).



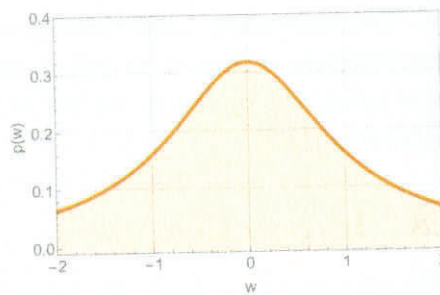
(a) Gaussian distribution



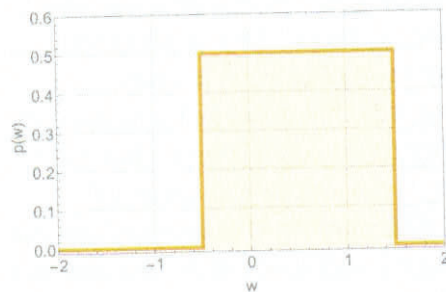
(b) Laplace distribution



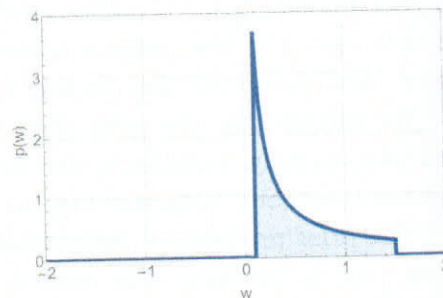
(c) Exponential distribution



(d) Cauchy distribution



(e) Uniform distribution



(f) Jeffreys distribution

Figure 7.1: Various probability distributions.

Chapter 8

Gaussian distr

... [W]e may imagine Chance
Competition with each other, ...

8.1 Introduction

The Gaussian distribution is ubiquitous. Its first publication in 1809 by the German mathematician Carl Friedrich Gauss (1777--1855) was the culmination of a long history. In Figure 6.1. However, a comparison with the French-English mathematician

The Gaussian distribution has several properties related to 1) the central limit theorem, 2) the product of two Gaussian distributions, 4) the Fourier transform, and 5) the maximum entropy principle under constraints.

Rewriting the univariate Gaussian

$$N(w | \mu, \sigma) = (2\pi)^{-\frac{1}{2}} (\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(w - \mu)^2\right)$$

It matches better with the upcoming multivariate Gaussian distribution. The σ^2 is commonly known as the variance.

The Gaussian distribution is also the maximum entropy distribution, which explains the frequent