Conjugate prior

In Bayesian probability theory, if the posterior distribution $p(\theta \mid x)$ is in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called **conjugate distributions,** and the prior is called a **conjugate prior** for the likelihood function $p(x \mid \theta)$.

A conjugate prior is an algebraic convenience, giving a closed-form expression for the posterior; otherwise numerical integration may be necessary. Further, conjugate priors may give intuition, by more transparently showing how a likelihood function updates a prior distribution.

The concept, as well as the term "conjugate prior", were introduced by Howard Raiffa and Robert Schlaifer in their work on Bayesian decision theory.[1] A similar concept had been discovered independently by George Alfred Barnard.[2]

Contents

Example

The form of the conjugate prior can generally be determined by inspection of the probability density or probability mass function of a distribution. For example, consider a random variable which consists of the number of successes s in n Bernoulli trials with unknown probability of success q in [0,1]. This random variable will follow the binomial distribution, with a probability mass function of the form

$$p(s) = \binom{n}{s} q^s (1-q)^{n-s}$$

The usual conjugate prior is the beta distribution with parameters (α, β):

$$p(q) = \frac{q^{\alpha-1}(1-q)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$$

where α and β are chosen to reflect any existing belief or information ($\alpha = 1$ and $\beta = 1$ would give a uniform distribution) and $B(\alpha, \beta)$ is the Beta function acting as a normalising constant.

In this context, α and β are called *hyperparameters* (parameters of the prior), to distinguish them from parameters of the underlying model (here q). It is a typical characteristic of conjugate priors that the dimensionality of the hyperparameters is one greater than that of the parameters of the original distribution. If all parameters are scalar values, then this means that there will be one more hyperparameter than parameter; but this also applies to vector-valued and matrix-valued parameters. (See the general article on the exponential family, and consider also the Wishart distribution, conjugate prior of the covariance matrix of a multivariate normal distribution, for an example where a large dimensionality is involved.)

If we then sample this random variable and get s successes and $f = n - s$ failures, we have

$$\begin{aligned}
P(s, f \mid q = x) &= \binom{s+f}{s} x^s (1-x)^f, \\
P(q = x) &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}, \\
P(q = x \mid s, f) &= \frac{P(s, f \mid x) P(x)}{\int P(s, f \mid y) P(y) dy} \\
&= \frac{\binom{s+f}{s} x^{s+\alpha-1}(1-x)^{f+\beta-1} / \mathrm{B}(\alpha,\beta)}{\int_{y=0}^{1} \left(\binom{s+f}{s} y^{s+\alpha-1}(1-y)^{f+\beta-1} / \mathrm{B}(\alpha,\beta) \right) dy} \\
&= \frac{x^{s+\alpha-1}(1-x)^{f+\beta-1}}{\mathrm{B}(s+\alpha, f+\beta)},
\end{aligned}$$

which is another Beta distribution with parameters ($\alpha + s$, $\beta + f$). This posterior distribution could then be used as the prior for more samples, with the hyperparameters simply adding each extra piece of information as it comes.

Interpretations

Pseudo-observations

It is often useful to think of the hyperparameters of a conjugate prior distribution as corresponding to having observed a certain number of *pseudo-observations* with properties specified by the parameters. For example, the values α and β of a beta distribution can be thought of as corresponding to $\alpha - 1$ successes and $\beta - 1$ failures if the posterior mode is used to choose an optimal parameter setting, or α successes and β failures if the posterior mean is used to choose an optimal parameter setting. In general, for nearly all conjugate prior distributions, the hyperparameters can be interpreted in terms of pseudo-observations. This can help both in providing an intuition behind the often messy update equations, as well as to help choose reasonable hyperparameters for a prior.

Analogy with eigenfunctions

Conjugate priors are analogous to eigenfunctions in operator theory, in that they are distributions on which the "conditioning operator" acts in a well-understood way, thinking of the process of changing from the prior to the posterior as an operator.

In both eigenfunctions and conjugate priors, there is a *finite-dimensional* space which is preserved by the operator: the output is of the same form (in the same space) as the input. This greatly simplifies the analysis, as it otherwise considers an infinite-dimensional space (space of all functions, space of all distributions).

However, the processes are only analogous, not identical: conditioning is not linear, as the space of distributions is not closed under linear combination, only convex combination, and the posterior is only of the same *form* as the prior, not a scalar multiple.

Just as one can easily analyze how a linear combination of eigenfunctions evolves under application of an operator (because, with respect to these functions, the operator is diagonalized), one can easily analyze how a convex combination of conjugate priors evolves under conditioning; this is called using a *hyperprior,* and corresponds to using a mixture density of conjugate priors, rather than a single conjugate prior.

Dynamical system

One can think of conditioning on conjugate priors as defining a kind of (discrete time) dynamical system: from a given set of hyperparameters, incoming data updates these hyperparameters, so one can see the change in hyperparameters as a kind of "time evolution" of the system, corresponding to "learning". Starting at different points yields different flows over time. This is again analogous with the dynamical system defined by a linear operator, but note that since different samples lead to different inference, this is not simply dependent on time, but rather on data over time. For related approaches, see Recursive Bayesian estimation and Data assimilation.

Practical example

Suppose a rental car service operates in your city. Drivers can drop off and pick up cars anywhere inside the city limits. You can find and rent cars using an app.

Suppose you wish to find the probability that you can find a rental car within a short distance of your home address at any time of day.

Over three days you look at the app and find the following number of cars within a short distance of your home address: $\mathbf{x} = [3, 4, 1]$

If we assume the data comes from a Poisson distribution, we can compute the maximum likelihood estimate of the parameters of the model which is $\lambda = \frac{3+4+1}{3} \approx 2.67$. Using this maximum likelihood estimate we can compute the probability that there will be at least one car available on a given day: $p(x > 0|\lambda \approx 2.67) = 1 - p(x = 0|\lambda \approx 2.67) = 1 - \frac{2.67^0 e^{-2.67}}{0!} \approx 0.93$

This is the Poisson distribution that is *the* most likely to have generated the observed data $\mathbf{x}$. But the data could also have come from another Poisson distribution, e.g. one with $\lambda = 3$, or $\lambda = 2$, etc. In fact there is an infinite number of poisson distributions that *could* have generated the observed data and with relatively few data points we should be quite uncertain about which exact poisson distribution generated this data. Intuitively we should instead take a weighted average of the probability of $p(x > 0|\lambda)$ for each of those Poisson distributions, weighted by how likely they each are, given the data we've observed $\mathbf{x}$.

Generally, this quantity is known as the posterior predictive distribution $p(x|\mathbf{x}) = \int_\theta p(x|\theta)p(\theta|\mathbf{x})d\theta$, where x is a new data point, $\mathbf{x}$ is the observed data and θ are the parameters of the model. Using Bayes' theorem we can expand $p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}$, therefore $p(x|\mathbf{x}) = \int_\theta p(x|\theta)\frac{p(\mathbf{x}|\theta)p(\theta)}{p(\mathbf{x})}d\theta$. Generally, this integral is hard to compute. However, if you choose a conjugate prior distribution $p(\theta)$, a closed form expression can be derived. This is the posterior predictive column in the tables below.

Returning to our example, if we pick the Gamma distribution as our prior distribution over the rate of the poisson distributions, then the posterior predictive is the negative binomial distribution as can be seen from the last column in the table below. The Gamma distribution is parameterized by two hyperparameters α, β which we have to choose. By looking at plots of the gamma distribution we pick $\alpha = \beta = 2$, which seems to be a reasonable prior for the average number of cars. The choice of prior hyperparameters is inherently subjective and based on prior knowledge.

Given the prior hyperparameters α and β we can compute the posterior hyperparameters $\alpha' = \alpha + \sum_i x_i = 2 + 3 + 4 + 1 = 10$ and $\beta' = \beta + n = 2 + 3 = 5$

Given the posterior hyperparameters we can finally compute the posterior predictive of $p(x > 0|\mathbf{x}) = 1 - p(x = 0|\mathbf{x}) = 1 - NB\left(0 \,|\, 10, \frac{1}{1+5}\right) \approx 0.84$

This much more conservative estimate reflect the uncertainty in the model parameters, which the posterior predictive takes into account.

Table of conjugate distributions

Let n denote the number of observations. In all cases below, the data is assumed to consist of n points $\boldsymbol{x_1, \ldots, x_n}$ (which will be random vectors in the multivariate cases).

If the likelihood function belongs to the exponential family, then a conjugate prior exists, often also in the exponential family; see Exponential family: Conjugate distributions.

When likelihood function is a discrete distribution

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters[note 1]	Interpretation of hyperparameters	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	$\boldsymbol{\alpha, \beta \in \mathbb{R}}$	$\boldsymbol{\alpha + \sum_{i=1}^{n} x_i,\ \beta + n - \sum_{i=1}^{n} x_i}$	$\boldsymbol{\alpha}$ successes, $\boldsymbol{\beta}$ failures[note 3]	$\boldsymbol{p(\tilde{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}}$
Binomial	p (probability)	Beta	$\boldsymbol{\alpha, \beta \in \mathbb{R}}$	$\boldsymbol{\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i}$	$\boldsymbol{\alpha}$ successes, $\boldsymbol{\beta}$ failures[note 3]	$\boldsymbol{\mathrm{BetaBin}(\tilde{x} \mid \alpha', \beta')}$ (beta-binomial)
Negative binomial with known failure number, r	p (probability)	Beta	$\boldsymbol{\alpha, \beta \in \mathbb{R}}$	$\boldsymbol{\alpha + rn,\ \beta + \sum_{i=1}^{n} x_i}$	$\boldsymbol{\alpha}$ total successes, $\boldsymbol{\beta}$ failures[note 3] (i.e., $\boldsymbol{\frac{\beta}{r}}$ experiments, assuming $\boldsymbol{r}$ stays fixed)	$\boldsymbol{\mathrm{BetaNegBin}(\tilde{x} \mid \alpha', \beta')}$ (beta-negative binomial)
Poisson	λ (rate)	Gamma	$\boldsymbol{k, \theta \in \mathbb{R}}$	$\boldsymbol{k + \sum_{i=1}^{n} x_i,\ \frac{\theta}{n\theta + 1}}$	$\boldsymbol{k}$ total occurrences in $\boldsymbol{\frac{1}{\theta}}$ intervals	$\boldsymbol{\mathrm{NB}\left(\tilde{x} \mid k', \frac{\theta'}{\theta' + 1}\right)}$ (negative binomial)
			$\boldsymbol{\alpha, \beta}$[note 4]	$\boldsymbol{\alpha + \sum_{i=1}^{n} x_i,\ \beta + n}$	$\boldsymbol{\alpha}$ total occurrences in $\boldsymbol{\beta}$ intervals	$\boldsymbol{\mathrm{NB}\left(\tilde{x} \mid \alpha', \frac{1}{1 + \beta'}\right)}$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha \in \mathbb{R}^k}$	$\boldsymbol{\alpha + (c_1, \ldots, c_k)}$, where $\boldsymbol{c_i}$ is the number of observations in category i	$\boldsymbol{\alpha_i}$ occurrences of category $\boldsymbol{i}$[note 3]	$\boldsymbol{p(\tilde{x} = i) = \frac{\alpha_i'}{\sum_i \alpha_i'} = \frac{\alpha_i + c_i}{\sum_i \alpha_i + n}}$
Multinomial	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha \in \mathbb{R}^k}$	$\boldsymbol{\alpha + \sum_{i=1}^{n} \mathbf{x}_i}$	$\boldsymbol{\alpha_i}$ occurrences of category $\boldsymbol{i}$[note 3]	$\boldsymbol{\mathrm{DirMult}(\tilde{\mathbf{x}} \mid \alpha')}$ (Dirichlet-multinomial)
Hypergeometric with known total population size, N	M (number of target members)	Beta-binomial[3]	$\boldsymbol{n = N, \alpha, \beta}$	$\boldsymbol{\alpha + \sum_{i=1}^{n} x_i,\ \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i}$	$\boldsymbol{\alpha}$ successes, $\boldsymbol{\beta}$ failures[note 3]	
Geometric	p_0 (probability)	Beta	$\boldsymbol{\alpha, \beta \in \mathbb{R}}$	$\boldsymbol{\alpha + n,\ \beta + \sum_{i=1}^{n} x_i}$	$\boldsymbol{\alpha}$ experiments, $\boldsymbol{\beta}$ total failures[note 3]	

When likelihood function is a continuous distribution

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters[note 1]	Interpretation of hyperparameters	
Normal with known variance σ^2	μ (mean)	Normal	$\mu_0,\ \sigma_0^2$	$\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right),\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean μ_0	$\mathcal{N}(\tilde{x}\vert\mu_0',$
Normal with known precision τ	μ (mean)	Normal	$\mu_0,\ \tau_0^{-1}$	$\frac{\tau_0\mu_0+\tau\sum_{i=1}^n x_i}{\tau_0+n\tau},\ (\tau_0+n\tau)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean μ_0	$\mathcal{N}\left(\tilde{x}\mid\mu\right.$
Normal with known mean μ	σ^2 (variance)	Inverse gamma	$\alpha,\ \beta$ [note 6]	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	variance was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\vert\mu,$
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	$\nu,\ \sigma_0^2$	$\nu+n,\ \frac{\nu\sigma_0^2+\sum_{i=1}^n (x_i-\mu)^2}{\nu+n}$	variance was estimated from ν observations with sample variance σ_0^2	$t_{\nu}(\tilde{x}\vert\mu,$
Normal with known mean μ	τ (precision)	Gamma	$\alpha,\ \beta$[note 4]	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	precision was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\mid\mu$
Normal[note 7]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0,\ \nu,\ \alpha,\ \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n},\ \nu+n,\ \alpha+\frac{n}{2},$ $\beta+\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ ▪ $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0; variance was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x}\mid\right.$
Normal	μ and τ Assuming exchangeability	Normal-gamma	$\mu_0,\ \nu,\ \alpha,\ \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n},\ \nu+n,\ \alpha+\frac{n}{2},$ $\beta+\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ ▪ $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0, and precision was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x}\mid\right.$
Multivariate normal with known covariance matrix $\boldsymbol{\Sigma}$	$\boldsymbol{\mu}$ (mean vector)	Multivariate normal	$\boldsymbol{\mu}_0,\ \boldsymbol{\Sigma}_0$	$\left(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0+n\boldsymbol{\Sigma}^{-1}\bar{\mathbf{x}}\right),$ $\left(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}$ ▪ $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions) $\boldsymbol{\Sigma}_0^{-1}$ and with sample mean $\boldsymbol{\mu}_0$	$\mathcal{N}(\tilde{\mathbf{x}}\mid\boldsymbol{\mu}$
Multivariate normal with known precision matrix $\boldsymbol{\Lambda}$	$\boldsymbol{\mu}$ (mean vector)	Multivariate normal	$\boldsymbol{\mu}_0,\ \boldsymbol{\Lambda}_0$	$\left(\boldsymbol{\Lambda}_0+n\boldsymbol{\Lambda}\right)^{-1}\left(\boldsymbol{\Lambda}_0\boldsymbol{\mu}_0+n\boldsymbol{\Lambda}\bar{\mathbf{x}}\right),\left(\boldsymbol{\Lambda}_0+n\boldsymbol{\Lambda}\right)$ ▪ $\bar{\mathbf{x}}$ is the sample mean	mean was estimated from observations with total precision (sum of all individual precisions) $\boldsymbol{\Lambda}_0$ and with sample mean $\boldsymbol{\mu}_0$	$\mathcal{N}\left(\tilde{\mathbf{x}}\mid\boldsymbol{\mu}\right.$
Multivariate normal with known mean $\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$ (covariance matrix)	Inverse-Wishart	$\nu,\ \boldsymbol{\Psi}$	$n+\nu,\ \boldsymbol{\Psi}+\sum_{i=1}^n(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T$	covariance matrix was estimated from ν observations with sum of pairwise deviation products $\boldsymbol{\Psi}$	$t_{\nu'-p+1}\left($
Multivariate normal with known mean $\boldsymbol{\mu}$	$\boldsymbol{\Lambda}$ (precision matrix)	Wishart	$\nu,\ \mathbf{V}$	$n+\nu,\ \left(\mathbf{V}^{-1}+\sum_{i=1}^n(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T\right)^{-1}$	covariance matrix was estimated from ν observations with sum of pairwise deviation products $\mathbf{V}^{-1}$	$t_{\nu'-p+1}\left($
Multivariate normal	$\boldsymbol{\mu}$ (mean vector) and $\boldsymbol{\Sigma}$ (covariance matrix)	normal-inverse-Wishart	$\boldsymbol{\mu}_0,\ \kappa_0,\ \nu_0,\ \boldsymbol{\Psi}$	$\frac{\kappa_0\boldsymbol{\mu}_0+n\bar{\mathbf{x}}}{\kappa_0+n},\ \kappa_0+n,\ \nu_0+n,$ $\boldsymbol{\Psi}+\mathbf{C}+\frac{\kappa_0 n}{\kappa_0+n}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^T$ ▪ $\bar{\mathbf{x}}$ is the sample mean ▪ $\mathbf{C}=\sum_{i=1}^n(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T$	mean was estimated from κ_0 observations with sample mean $\boldsymbol{\mu}_0$; covariance matrix was estimated from ν_0 observations with sample mean $\boldsymbol{\mu}_0$ and with sum of pairwise deviation products $\boldsymbol{\Psi}=\nu_0\boldsymbol{\Sigma}_0$	$t_{\nu_0'-p+1}$
Multivariate normal	$\boldsymbol{\mu}$ (mean vector) and $\boldsymbol{\Lambda}$ (precision matrix)	normal-Wishart	$\boldsymbol{\mu}_0,\ \kappa_0,\ \nu_0,\ \mathbf{V}$	$\frac{\kappa_0\boldsymbol{\mu}_0+n\bar{\mathbf{x}}}{\kappa_0+n},\ \kappa_0+n,\ \nu_0+n,$ $\left(\mathbf{V}^{-1}+\mathbf{C}+\frac{\kappa_0 n}{\kappa_0+n}(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)(\bar{\mathbf{x}}-\boldsymbol{\mu}_0)^T\right)^{-1}$ ▪ $\bar{\mathbf{x}}$ is the sample mean ▪ $\mathbf{C}=\sum_{i=1}^n(\mathbf{x}_i-\bar{\mathbf{x}})(\mathbf{x}_i-\bar{\mathbf{x}})^T$	mean was estimated from κ_0 observations with sample mean $\boldsymbol{\mu}_0$; covariance matrix was estimated from ν_0 observations with sample mean $\boldsymbol{\mu}_0$ and with sum of pairwise deviation products $\mathbf{V}^{-1}$	$t_{\nu_0'-p+1}$ [4]
Uniform	$U(0,\theta)$	Pareto	$x_m,\ k$	$\max\{x_1,\ldots,x_n,x_m\},\ k+n$	k observations with maximum value x_m	
Pareto with known minimum x_m	k (shape)	Gamma	$\alpha,\ \beta$	$\alpha+n,\ \beta+\sum_{i=1}^n \ln\frac{x_i}{x_m}$	α observations with sum β of the order of magnitude of each observation (i.e. the logarithm of the ratio of each observation to the minimum x_m)	
Weibull with known shape β	θ (scale)	Inverse gamma[3]	a,b	$a+n,\ b+\sum_{i=1}^n x_i^\beta$	a observations with sum b of the βth power of each observation	
Log-normal	Same as for the normal distribution after applying the natural logarithm to the data for the posterior hyperparameters. Please refer to Fink (1997, pp. 21–22) to see					
Exponential	λ (rate)	Gamma	$\alpha,\ \beta$[note 4]		$\alpha-1$ observations that sum to β [5]	$\mathbf{Lomax}($

				$\alpha + n,\ \beta + \sum_{i=1}^{n} x_i$		(Lomax c
Gamma with known shape α	β (rate)	Gamma	$\alpha_0,\ \beta_0$	$\alpha_0 + n\alpha,\ \beta_0 + \sum_{i=1}^{n} x_i$	α_0/α observations with sum β_0	$\mathrm{CG}(\tilde{x} \mid$ [note 8]
Inverse Gamma with known shape α	β (inverse scale)	Gamma	$\alpha_0,\ \beta_0$	$\alpha_0 + n\alpha,\ \beta_0 + \sum_{i=1}^{n} \frac{1}{x_i}$	α_0/α observations with sum β_0	
Gamma with known rate β	α (shape)	$\propto \frac{a^{\alpha-1}\beta^{\alpha c}}{\Gamma(\alpha)^b}$	$a,\ b,\ c$	$a\prod_{i=1}^{n} x_i,\ b+n,\ c+n$	b or c observations (b for estimating α, c for estimating β) with product a	
Gamma[3]	α (shape), β (inverse scale)	$\propto \frac{p^{\alpha-1}e^{-\beta q}}{\Gamma(\alpha)^r \beta^{-\alpha s}}$	$p,\ q,\ r,\ s$	$p\prod_{i=1}^{n} x_i,\ q+\sum_{i=1}^{n} x_i,\ r+n,\ s+n$	α was estimated from r observations with product p; β was estimated from s observations with sum q	
Beta	α, β	$\propto \frac{\Gamma(\alpha+\beta)^k p^\alpha q^\beta}{\Gamma(\alpha)^k \Gamma(\beta)^k}$	$p,\ q,\ k$	$p\prod_{i=1}^{n} x_i,\ q\prod_{i=1}^{n}(1-x_i),\ k+n$	α and β were estimated from k observations with product p and product of the complements q	

See also

- Beta-binomial distribution

Notes

1. Denoted by the same symbols as the prior hyperparameters with primes added ('). For instance α is denoted α'
2. This is the posterior predictive distribution of a new data point $\tilde{x}$ given the observed data points, with the parameters marginalized out. Variables with primes indicate the posterior values of the parameters.
3. The exact interpretation of the parameters of a beta distribution in terms of number of successes and failures depends on what function is used to extract a point estimate from the distribution. The mean of a beta distribution is $\frac{\alpha}{\alpha+\beta}$, which corresponds to α successes and β failures, while the mode is $\frac{\alpha-1}{\alpha+\beta-2}$, which corresponds to $\alpha-1$ successes and $\beta-1$ failures. Bayesians generally prefer to use the posterior mean rather than the posterior mode as a point estimate, justified by a quadratic loss function, and the use of α and β is more convenient mathematically, while the use of $\alpha-1$ and $\beta-1$ has the advantage that a uniform $\mathrm{Beta}(1,1)$ prior corresponds to 0 successes and 0 failures. The same issues apply to the Dirichlet distribution.
4. β is rate or inverse scale. In parameterization of gamma distribution, $\theta = 1/\beta$ and $k = \alpha$.
5. This is the posterior predictive distribution of a new data point $\tilde{x}$ given the observed data points, with the parameters marginalized out. Variables with primes indicate the posterior values of the parameters. $\mathcal{N}$ and t_n refer to the normal distribution and Student's t-distribution, respectively, or to the multivariate normal distribution and multivariate t-distribution in the multivariate cases.
6. In terms of the inverse gamma, β is a scale parameter
7. A different conjugate prior for unknown mean and variance, but with a fixed, linear relationship between them, is found in the normal variance-mean mixture, with the generalized inverse Gaussian as conjugate mixing distribution.
8. $\mathrm{CG}()$ is a compound gamma distribution; $\beta'()$ here is a generalized beta prime distribution.

References

1. Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory*. Division of Research, Graduate School of Business Administration, Harvard University, 1961.
2. Jeff Miller et al. Earliest Known Uses of Some of the Words of Mathematics (http://jeff560.tripod.com/mathword.html), "conjugate prior distributions" (http://jeff560.tripod.com/c.html). Electronic document, revision of November 13, 2005, retrieved December 2, 2005.
3. Fink, Daniel (1997). "A Compendium of Conjugate Priors" (https://web.archive.org/web/20090529203101/http://www.people.cornell.edu/pages/df36/CONJINTRnew%20TEX.pdf) (PDF). CiteSeerX 10.1.1.157.5540 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.157.5540). Archived from the original (https://courses.physics.ucsd.edu/2018/Fall/physics210b/REFERENCES/conjugate_priors.pdf) (PDF) on May 29, 2009.
4. Murphy, Kevin P. (2007), *Conjugate Bayesian analysis of the Gaussian distribution* (http://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf) (PDF)
5. Liu, Han; Wasserman, Larry (2014). *Statistical Machine Learning* (https://www.stat.cmu.edu/~larry/=sml/Bayes.pdf#page=16) (PDF). p. 314.