Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search Wikipedia

Wikipedia
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
فارسی
Italiano
Русский
中文
Edit links

Conjugate prior

From Wikipedia, the free encyclopedia

In Bayesian probability theory, if the posterior distributions $p(\theta \mid x)$ are 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. For example, the Gaussian family is conjugate to itself (or *self-conjugate*) with respect to a Gaussian likelihood function: if the likelihood function is Gaussian, choosing a Gaussian prior over the mean will ensure that the posterior distribution is also Gaussian. This means that the Gaussian distribution is a conjugate prior for the likelihood that is also Gaussian. 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]

Consider the general problem of inferring a (continuous) distribution for a parameter θ given some datum or data *x*. From Bayes' theorem, the posterior distribution is equal to the product of the likelihood function $\theta \mapsto p(x \mid \theta)$ and prior $p(\theta)$, normalized (divided) by the probability of the data $p(x)$:

$$\begin{aligned} p(\theta \mid x) &= \frac{p(x \mid \theta)\,p(\theta)}{p(x)} \\ &= \frac{p(x \mid \theta)\,p(\theta)}{\int p(x \mid \theta')\,p(\theta')\,d\theta'} \end{aligned}$$

Let the likelihood function be considered fixed; the likelihood function is usually well-determined from a statement of the data-generating process[*example needed*]. It is clear that different choices of the prior distribution $p(\theta)$ may make the integral more or less difficult to calculate, and the product $p(x|\theta) \times p(\theta)$ may take one algebraic form or another. For certain choices of the prior, the posterior has the same algebraic form as the prior (generally with different parameter values). Such a choice is a *conjugate prior*.

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.

All members of the exponential family have conjugate priors.[3]

Part of a series on Statistics
Bayesian statistics

Theory
Admissible decision rule · Bayesian efficiency · Bayesian probability · Probability interpretations · Bayes' theorem · Bayes factor · Bayesian inference · Bayesian network · Prior · Posterior · Likelihood · **Conjugate prior** · Posterior predictive · Hyperparameter · Hyperprior · Principle of indifference · Principle of maximum entropy · Empirical Bayes method · Cromwell's rule · Bernstein–von Mises theorem · Schwarz criterion · Credible interval · Maximum a posteriori estimation · Radical probabilism

Techniques
Bayesian linear regression · Bayesian estimator · Approximate Bayesian computation · Markov chain Monte Carlo

Mathematics portal

V · T · E

Contents [hide]

Example [edit]

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 (α = 1 and β = 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* failures, we have

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

which is another Beta distribution with parameters (α + *s*, β + *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.

Pseudo-observations [edit]

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.

Interpretations [edit]

Analogy with eigenfunctions[*citation needed*] [edit]

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 [edit]

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.

Table of conjugate distributions [edit]

Let *n* denote the number of observations. In all cases below, the data is assumed to consist of *n* points $x_1, \ldots, x_n$ (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 [edit]

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

When likelihood function is a continuous distribution [edit]

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters	Posterior predictive[note 4]
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	$\mathcal{N}(\tilde{x} \mid \mu_0',{\sigma_0^2}'+\sigma^2)$[5]
Normal with known precision τ	μ (mean)	Normal	$\mu_0,\,\tau_0$	$\frac{\tau_0\mu_0+\tau\sum_{i=1}^n x_i}{\tau_0+n\tau},\,\tau_0+n\tau$	mean was estimated from observations with total precision (sum of all individual precisions)τ_0 and with sample mean μ_0	$\mathcal{N}\left(\tilde{x} \mid \mu_0',\frac{1}{\tau_0'}+\frac{1}{\tau}\right)$[5]
Normal with known mean μ	σ^2 (variance)	Inverse gamma	$\alpha,\,\beta$ [note 5]	$\alpha+\frac{n}{2},\,\beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	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)	$t_{2\alpha'}(\tilde{x} \mid \mu,\sigma^2=\beta'/\alpha')$[5]
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	$t_{\nu'}(\tilde{x} \mid \mu,{\sigma_0^2}')$[5]
Normal with known mean μ	τ (precision)	Gamma	$\alpha,\,\beta$[note 3]	$\alpha+\frac{n}{2},\,\beta+\frac{\sum_{i=1}^n (x_i-\mu)^2}{2}$	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)	$t_{2\alpha'}(\tilde{x} \mid \mu,\sigma^2=\beta'/\alpha')$[5]
Normal[note 6]	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	$t_{2\alpha'}\left(\tilde{x} \mid \mu',\frac{\beta'(\nu'+1)}{\nu'\alpha'}\right)$[5]
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	$t_{2\alpha'}\left(\tilde{x} \mid \mu',\frac{\beta'(\nu'+1)}{\alpha'\nu'}\right)$[5]
Multivariate normal with known covariance matrix **Σ**	**μ** (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 $\boldsymbol\mu_0$	$\mathcal{N}(\tilde{\mathbf{x}} \mid {\boldsymbol\mu_0}',{\boldsymbol\Sigma_0}'+\boldsymbol\Sigma)$[5]
Multivariate normal with known precision matrix **Λ**	**μ** (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 $\boldsymbol\mu_0$	$\mathcal{N}\left(\tilde{\mathbf{x}} \mid {\boldsymbol\mu_0}',({\boldsymbol\Lambda_0}'^{-1}+\boldsymbol\Lambda^{-1})^{-1}\right)$[5]
Multivariate normal with known mean **μ**	**Σ** (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 $\boldsymbol\Psi$	$t_{\nu'-p+1}\left(\tilde{\mathbf{x}} \mid \boldsymbol\mu,\frac{1}{\nu'-p+1}\boldsymbol\Psi'\right)$[5]
Multivariate normal with known mean **μ**	**Λ** (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 $\mathbf{V}^{-1}$	$t_{\nu'-p+1}\left(\tilde{\mathbf{x}} \mid \boldsymbol\mu,\frac{1}{\nu'-p+1}\mathbf{V}'^{-1}\right)$[5]
Multivariate normal	**μ** (mean vector) and **Σ** (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 $\boldsymbol\Psi=\nu_0\boldsymbol\Sigma_0$	$t_{{\nu_0}'-p+1}\left(\tilde{\mathbf{x}} \mid {\boldsymbol\mu_0}',\frac{{\kappa_0}'+1}{{\kappa_0}'({\nu_0}'-p+1)}\boldsymbol\Psi'\right)$[5]
Multivariate normal	**μ** (mean vector) and **Λ** (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 $\mathbf{V}^{-1}$	$t_{{\nu_0}'-p+1}\left(\tilde{\mathbf{x}} \mid {\boldsymbol\mu_0}',\frac{{\kappa_0}'+1}{{\kappa_0}'({\nu_0}'-p+1)}\mathbf{V}'^{-1}\right)$[5]
Uniform	$U(0,\theta)$	Pareto	$x_m,\,k$	$\max\{\,x_1,\ldots,x_n,x_\mathrm{m}\},\,k+n$	k observations with maximum value x_m	
Pareto with known minimum x_m	*k* (shape)	Gamma	$\alpha,\,\beta$	$\alpha+n,\,\beta+\sum_{i=1}^n \ln\frac{x_i}{x_\mathrm{m}}$	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[4]	a,b	$a+n,\,b+\sum_{i=1}^n x_i^\beta$	a observations with sum b of the β'th power of each observation	
Log-normal	Same as for the normal distribution after exponentiating the data					
Exponential	λ (rate)	Gamma	$\alpha,\,\beta$[note 3]	$\alpha+n,\,\beta+\sum_{i=1}^n x_i$	α observations that sum to β [6]	$\operatorname{Lomax}(\tilde{x} \mid \beta',\alpha')$ (Lomax distribution)
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	$\operatorname{CG}(\tilde{\mathbf{x}} \mid \alpha,{\alpha_0}',{\beta_0}')=\operatorname{\beta'}(\tilde{\mathbf{x}} \mid \alpha,{\alpha_0}',1,{\beta_0}')$ [note 7]
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 [4]	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	

See also [edit]

- Beta-binomial distribution

Notes [edit]

1. ^ *a b c d e f g h* 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 mode of a beta distribution is $\frac{\alpha-1}{\alpha+\beta-2}$, which corresponds to $\alpha-1$ successes and $\beta-1$ failures; but the mean is $\frac{\alpha}{\alpha+\beta}$, which corresponds to α successes and β failures. 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, but the use of α and β is somewhat more convenient mathematically and also corresponds well with the fact that Bayesians generally prefer to use the posterior mean rather than the posterior mode as a point estimate. The same issues apply to the Dirichlet distribution.
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. ^ *a b c* β is rate or inverse scale. In parameterization of gamma distribution, $\theta = 1/\beta$ and $k = \alpha$.
4. ^ 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.
5. ^ In terms of the inverse gamma, β is a scale parameter
6. ^ 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.
7. ^ $\mathrm{CG}()$ is a compound gamma distribution; $\beta'()$ here is a generalized beta prime distribution.

References [edit]

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, "conjugate prior distributions". Electronic document, revision of November 13, 2005, retrieved December 2, 2005.
3. ^ For a catalog, see Gelman, Andrew; Carlin, John B.; Stern, Hal S.; Rubin, Donald B. (2003). *Bayesian Data Analysis* (2nd ed.). CRC Press. ISBN 1-58488-388-X.
4. ^ *a b c* Fink, D. (1997). "A Compendium of Conjugate Priors". *DOE Contract 95-831* ((Caution: Unreliable source) In progress report: Beware of some errors in multivariate normal and models and Arethya's prior (see addendum)) `|format=` requires `|url=` (help). CiteSeerX 10.1.1.157.5540.
5. ^ *a b c d e f g h i j k l m* Murphy, Kevin P. (2007). "Conjugate Bayesian analysis of the Gaussian distribution" (PDF).
6. ^ Statistical Machine Learning, by Han Liu and Larry Wasserman, 2014, pg. 314: http://www.stat.cmu.edu/~larry/=sml/Bayes.pdf

Categories: Bayesian statistics | Conjugate prior distributions