More on Bayes and conjugate forms

Saad Mneimneh

1 A cool function, $\Gamma(x)$ (Gamma)

The Gamma function is defined as follows:

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$$

For $x > 1$, if we integrate by parts ($\int udv = uv - \int vdu$), we have:

$$\begin{aligned}\Gamma(x) &= -t^{x-1}e^{-t}|_0^\infty - \int_0^\infty -(x-1)t^{x-2}e^{-t}dt \\ &= 0 + (x-1)\int_0^\infty t^{x-2}e^{-t}dt \\ &= (x-1)\Gamma(x-1)\end{aligned}$$

Note also that $\Gamma(1) = \int_0^\infty e^{-t}dt = 1$. We conclude that if $x \geq 1$ is an integer, $\Gamma(x) = (x-1)!$. Therefore, the Gamma function represents a generalization of the factorial function defined only on the non-negative integers. Given any $a > 0$, the product of the following k terms can now be expressed as:

$$a \cdot (a+1) \cdot (a+2) \cdot \ldots \cdot (a+k-1) = \frac{\Gamma(a+k)}{\Gamma(a)}$$

Perhaps the most famous value of the Gamma function for a non-integer is $\Gamma(1/2) = \sqrt{\pi}$.

The Gamma function is a component of various probability density functions as we will see later on.

2 Chi-squared density

Let $X_1 \ldots X_n$ be independent normal variables such that $X_i \sim N(0,1)$, and consider the sum $V = X_1^2 + \ldots X_n^2$. What is the probability density of V?

For simplicity, let us first stop the sum at X_1^2, i.e. the probability density of X^2 if $X \sim N(0,1)$.

$$P(-\sqrt{y} - \delta \leq X \leq -\sqrt{y}) + P(\sqrt{y} \leq X \leq \sqrt{y} + \delta) = P(y \leq X^2 \leq (\sqrt{y} + \delta)^2)$$

$$f_X(-\sqrt{y})\delta + f_X(\sqrt{y})\delta = f_{X^2}(y)(\delta^2 + 2\sqrt{y}\delta)$$

Taking the limit as $\delta \to 0$, we have:

$$f_{X^2}(y) = \frac{\phi(-\sqrt{y}) + \phi(\sqrt{y})}{2\sqrt{y}} = \frac{1}{2\sqrt{\pi}}(\frac{y}{2})^{\frac{1}{2}-1}e^{-\frac{y}{2}}$$

The expression above is arranged to reveal a form similar to the integrand of the Gamma function. Therefore, using a change of variable $t = y/2$:

$$\int_0^\infty (\frac{y}{2})^{\frac{1}{2}-1}e^{-\frac{y}{2}}dy = 2\int_0^\infty t^{\frac{1}{2}-1}e^{-t}dt = 2\Gamma(\frac{1}{2}) = 2\sqrt{\pi}$$

This shows that the density integrates to 1. Another way for obtaining the above density is to use an appropriate change of variable. To illustrate this general technique, assume that $f_X(x)$ is given, and that we are interested in finding $f_Y(y)$ where $y = g(x)$. Observe that $dy = g'(x)dx$ and, therefore, $dx = dy/g'(x)$. Hence,

$$\int f_X(x)dx = \int \frac{f_X(x)}{|g'(x)|}dy = 1$$

with appropriate adjustment of the integral range. If $x = g^{-1}(y)$ is uniquely obtained from y, then we can write the above as follows:

$$\int \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}dy = 1$$

implying that

$$f_Y(y) = \frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))}$$

If, however, x is not uniquely obtained from y, then we add up the contribution of all solutions. Let us apply this to our example where $f_X(x) = \phi(x)$ and $y = g(x) = x^2$. We have $g'(x) = 2x$ and $x = \pm\sqrt{y}$. Therefore, $|g'(x)| = 2\sqrt{y}$.

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}$$

and since $f_X(x) = \phi(x)$ is symmetric, we get

$$f_Y(y) = \frac{2\phi(\sqrt{y})}{2\sqrt{y}} = \frac{\phi(\sqrt{y})}{\sqrt{y}} = \frac{1}{2\sqrt{\pi}}(\frac{y}{2})^{\frac{1}{2}-1}e^{-\frac{y}{2}}$$

which is as obtained before.

The above suggests also that we can generalize the form of the density for $k \geq 1$ as follows:

$$\frac{1}{2\Gamma(\frac{k}{2})}(\frac{y}{2})^{\frac{k}{2}-1}e^{-\frac{y}{2}}$$

where $f_{X^2}(y)$ being the special case when $k = 1$. This is called the χ^2 (Chi-squared) density with parameter k, or k degrees of freedom. It turns out χ^2 is exactly the density for V when $k = n$.

$$V = X_1^2 + \ldots + X_n^2 \sim \chi_n^2$$

This is because the sum of two χ^2 independent random variables with parameters k_1 and k_2 is a χ^2 random variable with parameter $k = k_1 + k_2$. To prove this, one can use the transform of a χ^2 random variable. Recall that the transform of a random variable Z is $E[e^{sZ}]$. Therefore, the transform of $Z_1 + Z_2$ is $E[e^{s(Z_1+Z_2)}] = E[e^{sZ_1+sZ_2}] = E[e^{sZ_1}e^{sZ_2}] = E[e^{sZ_1}]E[e^{sZ_2}]$ if Z_1 and Z_2 are independent. It is easy to show that the transform of a χ_k^2 random variable is $(1-2s)^{-k/2}$ and, therefore, the transform of the sum of two independent χ^2 random variables with parameters k_1 and k_2 is $(1-2s)^{-k_1/2}(1-2s)^{-k_2/2} = (1-2s)^{-(k_1+k_2)/2}$, which is the transform of a χ_k^2 random variable where $k = k_1 + k_2$.

Example: Fairness of dice. Consider throwing a pair of dice, and let S be the random variable corresponding to a given total on a single throw. If the pair of dice is fair, S can take the following values with the given probabilities:

2	3	4	5	6	7	8	9	10	11	12
$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

Assume that we throw the pair of dice $n = 144$ times, and let Y_s be the random variable corresponding to the number of times we obtain a particular value s for S. Note that Y_s is the sum of n Bernoulli trials with success probability p_s as given in the above table. Consider the following real data:

s	2	3	4	5	6	7	8	9	10	11	12
y_s	2	4	10	12	22	29	21	15	14	9	6
np_s	4	8	12	16	20	24	20	16	12	8	4

How can we test whether or not the given pair of dice is loaded? For $s = 2, \ldots 12$, and in general for $s = 1 \ldots k$, consider the following random variable:

$$Z_s = \frac{Y_s - np_s}{\sqrt{np_s(1-p_s)}}$$

where Y_s is the number of times outcome s occurs, and p_s is the corresponding probability (thus np_s is the expected number of times s should occur).

A natural way is to consider $\sum_i Z_i^2$ to see whether this sum is (probabilistically) too high or too low. By the central limit theorem, $Z_s \sim N(0,1)$ when n is large. Therefore, $\sum_s Z_s^2 \sim \chi_k^2$. But the Zs are not completely independent! For instance, Y_k can be computed if $Y_1, \ldots, Y_{k-1}$ are known (because $\sum_s Y_s = n$).

Instead, let us consider

$$Z_s^* = \frac{Y_s - np_s}{\sqrt{np_s}}$$

and, for simplicity, assume we only have two possible outcomes ($Y_1 + Y_2 = n, p_1 + p_2 = 1$).

$$Z_1^{*^2} + Z_2^{*^2} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(Y_2 - np_2)^2}{np_2} = \frac{(Y_1 - np_1)^2}{np_1} + \frac{(-Y_1 + np_1)^2}{n(1 - p_1)}$$

$$= \frac{(Y_1 - np_1)^2}{np_1(1 - p_1)} = Z_1^2$$

We know that $Z_1^2 \sim \chi_1^2$. In general (proof omitted), if $k - d$ variables are sufficient to determine all k, then

$$V = \sum_{i=1}^{k} Z_i^{*^2} \sim \chi_{k-d}^2$$

Applying this to our dice problem ($k = 11, d = 1$), we compute $V = 7.14583$. From the table for χ_{10}^2, we can find $P(V \leq 7.14583) = \int_0^V \chi_{10}^2(y)dy$, and see that $V = 7.14583$ falls within the entries for 25% and 50%, so it is not significantly high or significantly low; thus the pair of dice is satisfactory (not loaded).

3 Chi-squared prior

To Keep a Bayesian spirit, the χ^2 density provides a conjugate prior in a number of cases. For instance, let $x_1, \ldots, x_n$ be independent Poisson random variables with parameter λ. Then

$$P(x_1, \ldots, x_n|\lambda) \propto \lambda^T e^{-n\lambda}$$

where $T = \sum_i x_i$.

If $m\lambda \sim \chi_k^2$ for a prior, i.e.

$$f(\lambda) = \frac{m}{2\Gamma(k/2)}(\frac{m\lambda}{2})^{k/2-1}e^{-\frac{m\lambda}{2}}$$

then

$$f(\lambda|x_1, \ldots, x_n) \propto \lambda^T e^{-n\lambda}(\frac{m\lambda}{2})^{k/2-1}e^{-\frac{m\lambda}{2}}$$

$$f(\lambda|x_1, \ldots, x_n) \propto (\frac{(2n+m)\lambda}{2})^{(k+2T)/2-1}e^{-\frac{(2n+m)\lambda}{2}}$$

$$(2n+m)\lambda|x_1, \ldots, x_n \sim \chi_{k+2T}^2$$

Now consider n independent Gaussian random variables with a variance σ^2 that is unknown (mean μ is known).

$$X_i|\sigma^2 \sim N(\mu, \sigma^2)$$

What would be an appropriate prior for σ^2 (a conjugate one)? Note that

$$f(\sigma^2|x_1,\ldots,x_n) \propto \frac{1}{\sigma^n} e^{-\frac{\sum_i (x_i-\mu)^2}{2\sigma^2}} f(\sigma^2)$$

Let $S = \sum_i (x_i - \mu)^2$, then:

$$f(\sigma^2|x_1,\ldots,x_n) \propto (1/\sigma^2)^{n/2} e^{-\frac{S/\sigma^2}{2}} f(\sigma^2)$$

Therefore,

$$f(\sigma^2|x_1,\ldots,x_n)|\frac{d\sigma^2}{d(1/\sigma^2)}| \propto (1/\sigma^2)^{n/2} e^{-\frac{S/\sigma^2}{2}} f(\sigma^2)|\frac{d\sigma^2}{d(1/\sigma^2)}|$$

The term $|d\sigma^2/d(1/\sigma^2)|$ adjusts for changing the variable from σ^2 to $1/\sigma^2$ so that integration of the density is with respect to $1/\sigma^2$ rather than σ^2. We get:

$$f(1/\sigma^2|x_1,\ldots,x_n) \propto (1/\sigma^2)^{n/2} e^{-\frac{S/\sigma^2}{2}} f(1/\sigma^2)$$

If $S_0/\sigma^2 \sim \chi^2_k$ for some S_0, then:

$$f(1/\sigma^2) \propto (1/\sigma^2)^{k/2-1} e^{-\frac{S_0/\sigma^2}{2}}$$

Therefore,

$$f(1/\sigma^2|x_1,\ldots,x_n) \propto (1/\sigma^2)^{\frac{n+k}{2}-1} e^{-\frac{(S+S_0)/\sigma^2}{2}}$$

$$(S+S_0)/\sigma^2|x_1,\ldots,x_n \sim \chi^2_{n+k}$$

Example: Consider the following sample ($n = 20$):

9	18	21	26	14
18	22	27	15	19
22	29	15	19	24
30	16	20	24	32

We can compute $\bar{x} \approx 21$, and let us, for simplicity, assume that this is the real value of μ. In this case, $S = 664$. Now one may argue that knowledge of k and S_0 is not available. In this case, let $k = 0$ and $S_0 = 0$ to get the following (improper) prior:

$$f(1/\sigma^2) \propto (1/\sigma^2)^{-1}$$

Therefore, the posterior will be given by

$$644/\sigma^2 \sim \chi^2_{20}$$

and from the tables we see that σ^2 is between 20 and 75 with probability 0.95.

$$f(\log\sigma^2) \propto (1/\sigma^2)^{-1}/|\frac{d(\log\sigma^2)}{d(1/\sigma^2)})| = (1/\sigma^2)^{-1}/|-\frac{d(\log(1/\sigma^2))}{d(1/\sigma^2)})| \propto 1$$

which is uniform in $\log\sigma^2 \in (-\infty, +\infty)$.

This prompts the following question: is the improper uniform prior equivalent to some conjugate form? In other words, is there a function $g(\theta)$ such that $f(\theta|x)$ has a valid density when $f(g(\theta)) \propto 1$? This is equivalent to the following statement (why?):

$$f(\theta|x) \propto f(x|\theta)|\frac{dg(\theta)}{d\theta}|$$

The answer to this of course depends on $f(x|\theta)$. For our example, it is clear that we need

$$|\frac{dg(1/\sigma^2)}{d(1/\sigma^2)}| = (1/\sigma^2)^{-1}$$

which is satisfied if $g(1/\sigma^2) = \log\sigma^2$. If we recall the example of deciding on the mean μ of independent Gaussian samples, $g(\mu) = \mu$ is appropriate.

Example: Let $f(x|\lambda) = \lambda e^{-\lambda x}$ (exponential density). Then

$$f(\lambda|x) \propto \lambda e^{-\lambda x}|\frac{dg(\lambda)}{d\lambda}|$$

Obviously, $g(\lambda) = \log\lambda$ provides a valid posterior density (exponential). Therefore, the improper uniform prior $f(\log\lambda) \propto 1$ works.