Introduction to Bayesian Analysis

Lecture Notes for EEB 596z, ©B. Walsh 2002

As opposed to the point estimators (means, variances) used by **classical statistics**, **Bayesian statistics** is concerned with generating the posterior distribution of the unknown parameters given both the data and some prior density for these parameters. As such, Bayesian statistics provides a much more complete picture of the uncertainty in the estimation of the unknown parameters, especially after the confounding effects of nuisance parameters are removed.

Our treatment here is intentionally quite brief and we refer the reader to Lee (1997) and Draper (2000) for a complete introduction to Bayesian analysis, and the introductory chapters of Tanner (1996) for a more condensed treatment. While very deep (and very subtle) differences in philosophy separate hard-core **Bayesians** from hard-core **frequentists** (Efron 1986, Glymour 1981), our treatment here of Bayesian methods is motivated simply by their use as a powerful statistical tool.

BAYES' THEOREM

The foundation of Bayesian statistics is **Bayes' theorem**. Suppose we observe a random variable y and wish to make inferences about another random variable θ, where θ is drawn from some distribution $p(\theta)$. From the definition of conditional probability,

$$\Pr(\theta \mid y) = \frac{\Pr(y, \theta)}{\Pr(y)} \tag{1a}$$

Again from the definition of conditional probability, we can express the joint probability by conditioning on θ to give

$$Pr(y, \theta) = \Pr(y \mid \theta) \Pr(\theta) \tag{1b}$$

Putting these together gives Bayes' theorem:

$$\Pr(\theta \mid y) = \frac{\Pr(y \mid \theta) \Pr(\theta)}{\Pr(y)} \tag{2a}$$

With n possible outcomes $(\theta_1, \cdots, \theta_n)$,

$$\Pr(\theta_j \mid y) = \frac{\Pr(y \mid \theta_j) \Pr(\theta_j)}{\Pr(y)} = \frac{\Pr(y \mid \theta_j)}{\sum_{i=1}^{n} \Pr(b_i)\ \Pr(y \mid \theta_i)} \tag{2b}$$

$\Pr(\theta)$ is the **prior distribution** of the possible θ values, while $\Pr(\theta \mid y)$ is the **posterior distribution** of θ given the observed data y. The origin of Bayes' theorem has a fascinating history (Stigler 1983). It is named after the Rev. Thomas Bayes, a priest who never published a mathematical paper in his lifetime. The paper in which the theorem appears was posthumously read before the Royal Society by his friend Richard Price in 1764. Stigler suggests it was first discovered by Nicholas Saunderson, a blind mathematician/optician who, at age 29, became Lucasian Professor of Mathematics at Cambridge (the position held earlier by Issac Newton).

Example 1. Suppose one in every 1000 families has a genetic disorder (sex-bias) in which they produce only female offspring. For any particular family we can define the (indicator) random variable

$$\theta = \begin{cases} 0 & \text{normal family} \\ 1 & \text{sex-bias family} \end{cases}$$

Suppose we observe a family with 5 girls and no boys. What is the probability that this family is a sex-bias family? From prior information, there is a 1/1000 chance that any randomly-chosen family is a sex-bias family, so $\Pr(\theta = 1) = 0.001$. Likewise $y =$ five girls, and Pr(five girls | sex bias family) = 1. This is $\Pr(y \mid \theta)$. It remains to compute the probability that a random family from the population with five children has all girls. Conditioning over all types of families (normal + sex-bias), Pr(5 girls) = Pr(5 girls | normal)*Pr(normal) + Pr(5 girls | sex-bias)*Pr(sex-bias), giving

$$\Pr(y) = (1/2)^5 \cdot (999/1000) + 1 \cdot (1/1000) = 0.0322$$

Hence,

$$\Pr(\theta = 1 \mid y = 5 \text{ girls}) = \frac{\Pr(y \mid \theta = 1)\Pr(\theta = 1)}{\Pr(y)} = \frac{1 \cdot 0.001}{0.0322} = 0.032$$

Thus, a family with five girls is 32 times more likely than a random family to have the sex-bias disorder.

Example 2. Suppose a major gene (with alleles **Q** and **q**) underlies a character of interest. The distribution of phenotypic values for each major locus genotype follows a normal distribution with variance one and means 2.1, 3.5, and 1.3 for **QQ**, **Qq**, and **qq** (respectively). Suppose the frequencies of these genotypes for a random individual drawn from the population are 0.3, 0.2, and 0.5 (again for **QQ**, **Qq**, and **qq** respectively). If an individual from this population has a phenotypic value of 3, what is the probability of it being **QQ**? **Qq**? **qq**?

Let $\varphi(z \mid \mu, 1) = (2\pi)^{-1/2} e^{-(z-\mu)^2/2}$ denote the density function for a normal with mean μ and variance one. To apply Bayes' theorem, the values for the priors and the conditionals are as follows:

Genotype, G	Pr(G)	Pr(y|G)	Pr(G)·Pr(y|G)
QQ	0.3	$\varphi(3 \mid 2.1, 1) = 0.177$	0.053
Qq	0.2	$\varphi(3 \mid 3.5, 1) = 0.311$	0.062
qq	0.5	$\varphi(3 \mid 1.3, 1) = 0.022$	0.011

Since $\sum_G$ Pr(G)·Pr(y | G) = 0.126, Bayes' theorem gives the posterior probabilities for the genotypes given the observed value of 3 as:

$$\Pr(\mathbf{QQ} \mid y = 3) = 0.177/0.126 = 0.421$$

$$\Pr(\mathbf{Qq} \mid y = 3) = 0.311/0.126 = 0.491$$

$$\Pr(\mathbf{qq} \mid y = 3) = 0.022/0.126 = 0.088$$

Thus, there is a 42 percent chance this individual has genotype **QQ**, a 49 percent chance it is **Qq**, and only an 8.8 percent chance it is **qq**.

Finally, the continuous multivariate version of Bayes' theorem is

$$p(\boldsymbol{\Theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{\int p(\mathbf{y}, \boldsymbol{\Theta})\, d\boldsymbol{\Theta}} \tag{3}$$

where $\boldsymbol{\Theta} = (\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(k)})$ is a vector of k (potentially) continuous variables. As with the univariate case, $p(\boldsymbol{\Theta})$ is the assumed prior distribution of the unknown parameters, while $p(\boldsymbol{\Theta} \mid \mathbf{y})$ is the posterior distribution given the prior $p(\boldsymbol{\Theta})$ and the data $\mathbf{y}$.

FROM LIKELIHOOD TO BAYESIAN ANALYSIS

The method of maximum likelihood and Bayesian analysis are closely related. Suppose $\ell(\boldsymbol{\Theta} \mid \mathbf{x})$ is the assumed likelihood function. Under ML estimation, we would compute the mode (the maximal value of ℓ, as a function of $\boldsymbol{\Theta}$ given the data $\mathbf{x}$) of the likelihood function, and use the local curvature to construct confidence intervals. Hypothesis testing follows using likelihood-ratio (LR) statistics. The strengths of ML estimation rely on its *large-sample* properties, namely that when the sample size is sufficiently large, we can assume both normality of the test statistic about its mean and that LR tests follow χ^2 distributions. These nice features don't necessarily hold for small samples.

An alternate way to proceed is to start with some initial knowledge/guess about the distribution of the unknown parameter(s), $p(\boldsymbol{\Theta})$. From Bayes' theorem, the data (likelihood) augment the prior distribution to produce a posterior distribution,

$$p(\boldsymbol{\Theta} \mid \mathbf{x}) = \frac{1}{p(\mathbf{x})} \cdot p(\mathbf{x} \mid \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}) \tag{4a}$$

$$= \begin{pmatrix} \text{normalizing} \\ \text{constant} \end{pmatrix} \cdot p(\mathbf{x} \mid \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}) \tag{4b}$$

$$= \text{constant} \cdot \text{likelihood} \cdot \text{prior} \tag{4c}$$

as $p(\mathbf{x} \mid \boldsymbol{\Theta}) = \ell(\boldsymbol{\Theta} \mid \mathbf{x})$ is just the likelihood function. $1/p(\mathbf{x})$ is a constant (with respect to $\boldsymbol{\Theta}$), because our concern is the distribution over θ. Because of this, the posterior distribution is often written as

$$p(\boldsymbol{\Theta} \mid \mathbf{x}) \propto \ell(\boldsymbol{\Theta} \mid \mathbf{x}) p(\boldsymbol{\Theta}) \tag{4d}$$

where the symbol $\propto$ means "proportional to" (equal up to a constant). Note that the constant $p(\mathbf{x})$ normalizes $p(\mathbf{x} \mid \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta})$ to one, and hence can be obtained by integration,

$$p(\mathbf{x}) = \int_{\boldsymbol{\Theta}} p(\mathbf{x} \mid \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}) d\boldsymbol{\Theta} \tag{5}$$

The dependence of the posterior on the prior (which can easily be assessed by trying different priors) provides an indication of how much information on the unknown parameter values is contained in the data. If the posterior is highly dependent on the prior, then the data likely has little signal, while if the posterior is largely unaffected under different priors, the data are likely highly informative. To see this, taking logs on Equation 4c (and ignoring the normalizing constant) gives

$$\log(\text{posterior}) = \log(\text{likelihood}) + \log(\text{prior}) \tag{6}$$

Marginal Posterior Distributions

Often, only a subset of the unknown parameters is really of concern to us, the rest being **nuisance parameters** that are really of no concern to us. A very strong feature of Bayesian analysis is that we can remove the effects of the nuisance parameters by simply integrating them out of the posterior distribution to generate a **marginal posterior distribution** for the parameters of interest. For example, suppose the mean and variance of data coming from a normal distribution are unknown, but our real interest is in the variance. Estimating the mean introduces additional uncertainly into our variance estimate. This is not fully capture in standard classical approaches, but under a Bayesian analysis, the posterior marginal

distribution for σ^2 is simply

$$p(\,\sigma^2 \,|\, \mathbf{x}) = \int p(\,\mu, \sigma^2 \,|\, \mathbf{x}\,)\, d\mu$$

The marginal posterior may involve several parameters (generating **joint marginal posteriors**). Write the vector of unknown parameters as $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_n)$, where $\boldsymbol{\Theta}_n$ is the vector of nuisance parameters. Integrating over $\boldsymbol{\Theta}_n$ gives the desired marginal as

$$p(\boldsymbol{\Theta}_1 \,|\, \mathbf{y}) = \int_{\boldsymbol{\Theta}_n} p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_n \,|\, \mathbf{y})\, d\boldsymbol{\Theta}_n \tag{7}$$

SUMMARIZING THE POSTERIOR DISTRIBUTION

How do we extract a Bayes estimator for some unknown parameter θ? If our mindset is to use some sort of point estimator (as is usually done in classical statistics), there are a number of candidates. We could follow maximum likelihood and use the **mode of the distribution** (its maximal value), with

$$\widehat{\theta} = \max_{\theta} [\, p(\,\theta \,|\, \mathbf{x}\,)] \tag{8a}$$

We could take the **expected value of** θ given the posterior,

$$\widehat{\theta} = E[\,\theta \,|\, \mathbf{x}\,] = \int \theta\, p(\,\theta \,|\, \mathbf{x}\,) d\theta \tag{8b}$$

Another candidate is the **medium of the posterior distribution**, where the estimator satisfies $\Pr(\theta > \widehat{\theta} \,|\, \mathbf{x}) = \Pr(\theta < \widehat{\theta} \,|\, \mathbf{x}) = 0.5$, hence

$$\int_{\widehat{\theta}}^{+\infty} p(\,\theta \,|\, \mathbf{x}\,) d\theta = \int_{-\infty}^{\widehat{\theta}} p(\,\theta \,|\, \mathbf{x}\,) d\theta = \frac{1}{2} \tag{8c}$$

However, using any of the above estimators, or even all three simultaneously, loses the full power of a Bayesian analysis, as ***the full estimator is the entire posterior density itself***. If we cannot obtain the full form of the posterior distribution, it may still be possible to obtain one of the three above estimators. However, as we will see later, we can generally obtain the posterior by simulation using Gibbs sampling, and hence the Bayes estimate of a parameter is frequently presented as a frequency histogram from (Gibbs) samples of the posterior distribution.

Highest Density Regions (HDRs)

Given the posterior distribution, construction of confidence intervals is obvious. For example, a $100(1-\alpha)$ confidence interval is given by any $(L_{\alpha/2}, H_{\alpha/2})$ satisfying

$$\int_{L_{\alpha/2}}^{H_{\alpha/2}} p(\theta \mid \mathbf{x})\, d\theta = 1 - \alpha$$

To reduce possible candidates, one typically uses **highest density regions**, or **HDRs**, where for a single parameter the HDR $100(1-\alpha)$ region(s) are the shortest intervals giving an area of $(1-\alpha)$. More generally, if multiple parameters are being estimated, the HDR region(s) are those with the smallest *volume* in the parameter space. HDRs are also referred to as **Bayesian confidence intervals** or **credible intervals**.

It is critical to note that there is a profound difference between a confidence interval (CI) from classical (frequentist) statistics and a Bayesian interval. The interpretation of a classical confidence interval is that is we repeat the experiment a large number of times, and construct CIs in the same fashion, that $(1-\alpha)$ of the time the confidence interval with enclose the (unknown) parameter. With a Bayesian HDR, there is a $(1-\alpha)$ probability that the interval contains the true value of the unknown parameter. Often the CI and Bayesian intervals have essentially the same value, but again the interpretational difference remains. The key point is that the Bayesian prior allows us to make direct probability statements about θ, while under classical statistics we can only make statements about the behavior of the statistic if we repeat an experiment a large number of times. Given the important conceptual difference between classical and Bayesian intervals, Bayesians often avoid using the term confidence interval.

Bayes Factors and Hypothesis Testing

In the classical hypothesis testing framework, we have two alternatives. The null hypothesis H_0 that the unknown parameter θ belongs to some set or interval Θ_0 ($\theta \in \Theta_0$), versus the alternative hypothesis H_1 that θ belongs to the alternative set Θ_1 ($\theta \in \Theta_1$). Θ_0 and Θ_1 contain no common elements ($\Theta_0 \cap \Theta_1 = \oslash$) and the union of Θ_0 and Θ_1 contains the entire space of values for θ (i.e., $\Theta_0 \cup \Theta_1 = \Theta$).

In the classical statistical framework of the frequentists, one uses the observed data to test the significant of a particular hypothesis, and (if possible) compute a p-value (the probability p of observing the given value of the test statistic if the null hypothesis is indeed correct). Hence, at first blush one would think that the idea of a hypothesis test is trivial in a Bayesian framework, as using the posterior distribution

$$\Pr(\theta > \theta_0) = \int_{\theta_0} p(\,\theta \mid \mathbf{x})\, d\theta \quad \text{and} \qquad \Pr(\theta_0 < \theta < \theta_1) = \int_{\theta_0}^{\theta_1} p(\,\theta \mid \mathbf{x})\, d\theta$$

The kicker with a Bayesian analysis is that we also have prior information and Bayesian hypothesis testing addresses whether, *given the data*, we are more or less

inclined towards the hypothesis than we initially were. For example, suppose for the prior distribution of θ is such that $\Pr(\theta > \theta_0) = 0.10$, while for the posterior distribution $\Pr(\theta > \theta_0) = 0.05$. The later is significant at the 5 percent level in a classical hypothesis testing framework, but the data only doubles our confidence in the alternative hypothesis relative to our belief based on prior information. If $\Pr(\theta > \theta_0) = 0.50$ for the prior, then a 5% posterior probability would greatly increase our confidence in the alternative hypothesis. Hence, the prior probabilities certainly influence hypothesis testing.

To formalize this idea, let

$$p_0 = \Pr(\theta \in \Theta_0 \,|\, \mathbf{x}), \qquad p_1 = \Pr(\theta \in \Theta_1 \,|\, \mathbf{x}) \tag{9a}$$

denote the probability, given the observed data $\mathbf{x}$, that θ is in the null (p_0) and alternative (p_1) hypothesis sets. Note that these are posterior probabilities. Since $\Theta_0 \cap \Theta_1 = \oslash$ and $\Theta_0 \cup \Theta_1 = \Theta$, it follows that $p_0 + p_1 = 1$. Likewise, for the prior probabilities we have

$$\pi_0 = \Pr(\theta \in \Theta_0\,), \qquad \pi_1 = \Pr(\theta \in \Theta_1\,) \tag{9b}$$

Thus the **prior odds** of H_0 versus H_1 are π_0/π_1, while the **posterior odds** are p_0/p_1.

The **Bayes factor** $\mathbf{B}_0$ in favor of H_0 versus H_1 is given by the ratio of the posterior odds divided by the prior odds,

$$B_0 = \frac{p_0/p_1}{\pi_0/\pi_1} = \frac{p_0\pi_1}{p_1\pi_0} \tag{10a}$$

The Bayes factor is loosely interpreted as the odds in favor of H_0 versus H_1 that are given by the data. Since $\pi_1 = 1 - \pi_0$ and $p_1 = 1 - p_0$, we can also express this as

$$B_0 = \frac{p_0(1-\pi_0)}{\pi_0(1-p_0)} \tag{10b}$$

Likewise, by symmetry note that the Bayes factor B_1 in favor of H_1 versus H_0 is just

$$B_1 = 1/B_0 \tag{10c}$$

Consider the first case where the prior and posterior probabilities for the null were 0.1 and 0.05 (respectively). The Bayes factor in favor of H_1 versus H_0 is given by

$$B_1 = \frac{\pi_0(1-p_0)}{p_0(1-\pi_0)} = \frac{0.5 \cdot 0.95}{0.05 \cdot 0.5} = 2.11$$

Similarly, for the second example where the prior for the null was 0.5,

$$B_1 = \frac{0.1 \cdot 0.95}{0.05 \cdot 0.5} = 19$$

When the hypotheses are simple, say $\Theta_0 = \theta_0$ and $\Theta_1 = \theta_1$, then for $i = 0, 1$,

$$p_i \propto p(\theta_i)\, p(\mathbf{x} \,|\, \theta_i) = \pi_i\, p(\mathbf{x} \,|\, \theta_i)$$

Thus

$$\frac{p_0}{p_1} = \frac{\pi_0\, p(\mathbf{x} \,|\, \theta_0)}{\pi_1\, p(\mathbf{x} \,|\, \theta_1)} \tag{11a}$$

and the Bayes factor (in favor of the null) reduces the

$$B_0 = \frac{p(\mathbf{x} \,|\, \theta_0)}{p(\mathbf{x} \,|\, \theta_1)} \tag{11b}$$

which is simply a *likelihood ratio*.

When the hypotheses are **composite** (containing multiple members), things are slightly more complicated. First note that the prior distribution of θ conditioned on H_0 vs. H_1 is

$$p_i(\theta) = p(\theta)/\pi_i \qquad \text{for} \quad i = 0, 1 \tag{12}$$

as the total probability $\theta \in \Theta_i = \pi_i$, so that dividing by π_i normalizes the distribution to integrate to one. Thus

$$\begin{aligned} p_i = \Pr(\theta \in \Theta_i \,|\, \mathbf{x}) &= \int_{\theta \in \Theta_i} p(\theta \,|\, \mathbf{x}) d\theta \\ &\propto \int_{\theta \in \Theta_i} p(\theta) p(\mathbf{x} \,|\, \theta) d\theta \\ &= \pi_i \int_{\theta \in \Theta_i} p(\mathbf{x} \,|\, \theta) p_i(\theta) d\theta \end{aligned} \tag{13}$$

where the second step follows from Bayes' theorem (Equation 4d) and the final step follows from Equation (12), as $\pi_i\, p_i(\theta) = p(\theta)$. The Bayes factor in favor the null hypothesis thus becomes

$$B_0 = \left(\frac{p_0}{\pi_0}\right)\left(\frac{\pi_1}{p_1}\right) = \frac{\int_{\theta \in \Theta_0} p(\mathbf{x} \,|\, \theta) p_0(\theta) d\theta}{\int_{\theta \in \Theta_1} p(\mathbf{x} \,|\, \theta) p_1(\theta) d\theta} \tag{14}$$

which is a ratio of the weighted likelihoods of Θ_0 and Θ_1.

A compromise between Bayesian and classical hypothesis testing was suggested by Lindley (1965). If the goal is to conduct a hypothesis test of the form H_0: $\theta = \theta_0$ vs. H_2: $\theta \neq \theta_0$ and we assume a diffuse prior, then a significance test of level α follows by obtaining a $100(1-\alpha)\%$ HDR for the posterior and rejecting the null hypothesis if and only if θ is *outside* of the HDR.

See Lee (1997) are further discussions on hypothesis testing (or lack thereof) in a Bayesian framework.

THE CHOICE OF A PRIOR

Obviously, a critical feature of any Bayesian analysis is the choice of a prior. The key here is that when the data have sufficient signal, even a bad prior will still not greatly influence the posterior. In a sense, this is an asymptotic property of Bayesian analysis in that all but pathological priors will be overcome by sufficient amounts of data. As mentioned above, one can check the impact of the prior by seeing how stable to posterior distribution is to different choices of priors. If the posterior is highly dependent on the prior, then the data (the likelihood function) may not contain sufficient information. However, if the posterior is relatively stable over a choice of priors, then the data indeed contain significant information.

The **location** of a parameter (mean or mode) and its **precision** (the reciprocal of the variance) of the prior is usually more critical than its actual shape in terms of conveying prior information. The shape (family) of the prior distribution is often chosen to facilitate calculation of the prior, especially through the use of **conjugate priors** that, for a given likelihood function, return a posterior in the same distribution family as the prior (i.e., a gamma prior returning a gamma posterior when the likelihood is Poisson). We will return to conjugate priors and the end of these notes, but we first discuss other standard approaches for construction of priors.

Diffuse Priors

One of the most common priors is the **flat**, **uninformative**, or **diffuse** prior where the prior is simply a constant,

$$p(\theta) = k = \frac{1}{b-a} \qquad \text{for} \qquad a \leq \theta \leq b \tag{15a}$$

This conveys that we have no a priori reason to favor any particular parameter value over another. With a flat prior, the posterior just a constant times the likelihood,

$$p(\theta \,|\, \mathbf{x}) = C\, \ell(\theta \,|\, \mathbf{x}) \tag{15b}$$

and we typically write that $p(\theta \,|\, \mathbf{x}) \propto \ell(\theta \,|\, \mathbf{x})$. In many cases, classical expressions from frequentist statistics are obtained by Bayesian analysis assuming a flat prior.

If the variable of interest ranges over $(0, \infty)$ or $(-\infty, +\infty)$, then strictly speaking a flat prior does not exist, as if the constant takes on any non-zero value, the integral does not exist. In such cases a flat prior (assuming $p(\theta \,|\, \mathbf{x}) \propto \ell(\theta \,|\, \mathbf{x})$) is referred to as an **improper prior**.

Sufficient Statistics and Data-Transformed Likelihoods

Suppose we can write the likelihood for a given parameter θ and data vector $\mathbf{x}$ as

$$\ell(\, \theta \,|\, \mathbf{x} \,) = g\,[\, \theta - t(\mathbf{x}) \,] \tag{16}$$

Here the likelihood is a function $\ell = g(z)$, where $z = \theta - t(\mathbf{x})$. If the likelihood is of this form, the data $\mathbf{x}$ only influences θ by a translation on the scale of the function g, i.e., from $g(z)$ to $g(z + a)$. Further, note that $t(\mathbf{x})$ is the only value of the data that appears, andwe call the function t a **sufficient statistic**. Other data sets with different values of $\mathbf{x}$, but the same value of the sufficient statistic $t(\mathbf{x})$, have the same likelihood.

When the likelihood can be placed in the form of Equation 16, a shift in the data gives rise to the same functional form of the likelihood function except for a shift in location, from $(\theta + t[\mathbf{x}_1])$ to $(\theta + t[\mathbf{x}_2])$. Hence, this is a *natural scale* upon which to measure likelihoods, and on such a scale, a flat/diffuse prior seems natural.

Example 3. Consider n independent samples from a normal with unknown mean μ and known variance σ^2. Here

$$\ell(\,\mu\,|\,\mathbf{x}\,) \propto \exp\left(\frac{-(\mu - \overline{x}\,)^2}{2(\sigma^2/n)}\right)$$

Note immediately that $\overline{x}$ is a sufficient statistic for the mean, so that different data sets with the same mean (for n draws) have the same likelihood function for the unknown mean μ. Further note that

$$g(z) = \exp\left(\frac{-z^2}{2(\sigma^2/n)}\right)$$

Hence, a flat prior for μ seems appropriate.

What is the natural scale for a likelihood function that does not satisfy Equation 16? Suppose that the likelihood function can be written in **data-translated format** as

$$\ell(\,\theta\,|\,\mathbf{x}\,) = g\,[\,h(\theta) - t(\mathbf{x})\,] \tag{17}$$

When the likelihood function has this format, the natural scale for the unknown parameter is $h(\theta)$. Hence, a prior of the form $p[\,h(\theta)\,] = \text{constant}$ (a flat prior on $h[\,\theta\,]$) is suggested. Using a change of variables to transform $p[\,h(\theta)\,]$ back onto the θ scale suggests a prior on θ of the form

$$p(\theta) \propto \left|\,\frac{\partial h(\theta)}{\partial \theta}\,\right| \tag{18}$$

Example 4. Suppose the likelihood function assumes data follow an exponential distribution,

$$\ell(\theta \,|\, x) = (1/\theta)\exp(-x/\theta)$$

To express this likelihood in a data-translated format, we will make use of the useful fact that we can multiply any likelihood function by a constant and still have a likelihood function. In particular, since the data $\mathbf{x}$ is known (and hence treated as a constant), we can multiply the likelihood function by any function of the data, e.g. $f(\mathbf{x})\,\ell(\boldsymbol{\Theta} \,|\, x) \propto \ell(\boldsymbol{\Theta} \,|\, x)$. In this example, we simply multiply the likelihood function by x to give

$$\ell(\theta \,|\, x) = (x/\theta)\exp(-x/\theta)$$

Noting that

$$x/\theta = \exp\left[\ln\left(\frac{x}{\theta}\right)\right] = \exp\left[\ln x - \ln\theta\right]$$

we can express the likelihood as

$$\ell(\theta \,|\, x) = \exp[\,(\ln x - \ln\theta) - \exp(\ln x - \ln\theta)\,]$$

Hence, in data-translated format the likelihood function becomes

$$g(y) = \exp[y - \exp(y)\,], \qquad t(x) = \ln x, \qquad g(\theta) = \ln\theta$$

The "natural scale" for θ in this likelihood function is thus $\ln\theta$, and a natural prior is $p(\,\ln\theta\,) = \text{constant}$, giving the prior as

$$p(\theta) \propto \left|\frac{\partial \ln\theta}{\partial\theta}\right| = \frac{1}{\theta}$$

The Jeffreys' Prior

Suppose we cannot easily find the natural scale on which the likelihood is in data-translated format, or that such a decomposition does not exist. Jeffreys (1961) proposed a general prior in such cases, based on the Fisher information I of the likelihood. Recall that

$$I(\theta \,|\, \mathbf{x}\,) = -E_x\left(\frac{\partial^2 \ln \ell(\theta \,|\, \mathbf{x}\,)}{\partial\,\theta^2}\right)$$

Jeffreys' rule (giving the **Jeffreys' Prior**) is to take as the prior

$$p(\theta) \propto \sqrt{I(\theta \,|\, \mathbf{x}\,)} \tag{19}$$

A full discussion, with derivation, can be found in Lee (1997, Section 3.3).

Example 5. Consider the likelihood for n independent draws from a binomial,

$$\ell(\theta \,|\, \mathbf{x}) = C\theta^x(1-\theta)^{n-x}$$

where the constant C does not involve θ. Taking logs gives

$$L(\theta \,|\, \mathbf{x}) = \ln\left[\, \ell(\theta \,|\, \mathbf{x})\,\right] = \ln C + x\ln\theta + (n-x)\ln(1-\theta)$$

Thus

$$\frac{\partial L(\theta \,|\, \mathbf{x})}{\partial\theta} = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

and likewise

$$\frac{\partial^2 L(\theta \,|\, \mathbf{x})}{\partial\theta^2} = -\frac{x}{\theta^2} - (-1)\cdot(-1)\frac{n-x}{(1-\theta)^2} = -\left(\frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2}\right)$$

Since $E[\,x\,] = n\theta$, we have

$$-E_x\left(\frac{\partial^2 \ln \ell(\theta \,|\, \mathbf{x}\,)}{\partial\,\theta^2}\right) = \frac{n\theta}{\theta^2} + \frac{n(1-\theta)}{(1-\theta)^2} = n\,\theta^{-1}(1-\theta)^{-1}$$

Hence, the Jeffreys' Prior becomes

$$p(\theta) \propto \sqrt{\theta^{-1}(1-\theta)^{-1}} \propto \theta^{-1/2}(1-\theta)^{-1/2}$$

which is a Beta Distribution (which we discuss later).

When there are multiple parameters, $\mathbf{I}$ is the Fisher Information matrix, the matrix of the expected second partials,

$$\mathbf{I}(\,\boldsymbol{\Theta} \,|\, \mathbf{x}\,)_{ij} = -E_x\left(\frac{\partial^2 \ln \ell(\boldsymbol{\Theta} \,|\, \mathbf{x}\,)}{\partial\,\theta_i \partial\,\theta_j}\right)$$

In this case, the Jeffreys' Prior becomes

$$p(\boldsymbol{\Theta}) \propto \sqrt{\det[\mathbf{I}(\theta \,|\, \mathbf{x}\,)\,]} \tag{20}$$

Example 6. Suppose our data consists of n independent draws from a normal distribution with unknown mean and variance, μ and σ^2. In earlier notes on maximum likelihood estimation, we showed that the information matrix in this case is

$$\mathbf{I} = n \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

Since the determinant of a diagonal matrix is the product of the diagonal elements, we have $\det(\mathbf{I}) \propto \sigma^{-6}$, giving the Jeffreys' Prior for μ and σ^2 as

$$p(\boldsymbol{\Theta}) \propto \sqrt{\sigma^{-6}} = \sigma^{-3}$$

Since the prior does not involve μ, we assume a flat prior for μ (i.e. $p(\mu) =$ constant). Note that the prior distributions of μ and σ^2 are independent, as

$$p(\mu, \theta) = \text{constant } \cdot \sigma^{-3} = p(\mu) \cdot p(\sigma^2)$$

POSTERIOR DISTRIBUTIONS UNDER NORMALITY ASSUMPTIONS

To introduce the basic ideas of Bayesian analysis, consider the case where data are drawn from a normal distribution, so that the likelihood function for the ith observation, x_i is

$$\ell(\mu, \sigma^2 \,|\, x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{21a}$$

The resulting full likelihood for all n data points is

$$\ell(\mu \,|\, \mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{21b}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu n\overline{x} + n\mu^2\right)\right] \tag{21c}$$

Known Variance and Unknown Mean

Assume the variance (σ^2) is known, while the mean μ is unknown. For a Bayesian analysis, it remains to specify the prior for μ, $p(\mu)$. Suppose we assume a Gaussian prior, $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$, so that

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \tag{22}$$

The mean and variance of the prior, μ_0 and σ_0^2 are referred to as **hyperparameters**.

One important trick we will use throughout when calculating the posterior distribution is to ignore terms that are constants with respect to the unknown parameters. Suppose $\mathbf{x}$ denotes the data and $\boldsymbol{\Theta}_1$ is a vector of *known* model parameters, while $\boldsymbol{\Theta}_2$ is a vector of unknown parameters. If we can write the posterior as

$$p(\boldsymbol{\Theta}_2 \,|\, \mathbf{x}, \boldsymbol{\Theta}_1) = f(\mathbf{x}, \boldsymbol{\Theta}_1) \cdot g(\mathbf{x}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \tag{23a}$$

then

$$p(\boldsymbol{\Theta}_2 \,|\, \mathbf{x}, \boldsymbol{\Theta}_1) \propto g(\mathbf{x}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \tag{23b}$$

With the prior given by Equation 22, we can express the resulting posterior distribution as

$$\begin{aligned} p(\mu \,|\, \mathbf{x}) &\propto \ell(\mu \,|\, \mathbf{x}\,) \cdot p(\mu) \\ &\propto \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2} - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} x_i^2 - 2\mu n\overline{x} + n\mu^2 \right] \right) \end{aligned} \tag{24a}$$

We can factor out additional terms not involving μ to give

$$p(\mu \,|\, \mathbf{x}) \propto \exp\left(-\frac{\mu^2}{2\sigma_0^2} + \frac{\mu\,\mu_0}{\sigma_0^2} + \frac{\mu n\overline{x}}{\sigma^2} - \frac{n\mu^2}{2\sigma^2} \right) \tag{24b}$$

Factoring in terms of μ, the term in the exponential becomes

$$-\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) + \mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2} \right) = -\frac{\mu^2}{\sigma_*^2} + \frac{2\mu\mu_*}{2\sigma_*^2} \tag{25a}$$

where

$$\sigma_*^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \quad \text{and} \quad \mu_* = \sigma_*^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2} \right) \tag{25b}$$

Finally, by completing the square, we have

$$p(\mu \,|\, \mathbf{x}) \propto \exp\left(-\frac{(\mu - \mu_*)^2}{2\sigma_*^2} + f(\mathbf{x}, \mu_0, \sigma^2, \sigma_0^2) \right) \tag{25c}$$

The posterior density function for μ thus becomes

$$p(\mu \,|\, \mathbf{x}) \propto \exp\left(-\frac{(\mu - \mu_*)^2}{2\sigma_*^2} \right) \tag{26a}$$

Recalling that the density function for $z \sim \mathrm{N}(\alpha, \beta)$ is

$$p(z) \propto \exp\left(-\frac{(z - \alpha)^2}{2\beta} \right) \tag{26b}$$

shows that the posterior density function for μ is a normal with mean μ_* and variance σ_*^2, e.g.,

$$\mu \,|\, (\mathbf{x}, \sigma^2) \sim \mathrm{N}\left(\mu_*, \sigma_*^2\right) \tag{26c}$$

Notice that the posterior density is in the same form as the prior. This occurred because the prior **conjugated** with the likelihood function – the product of the prior and likelihood returned a distribution in the same family as the prior. The use of such **conjugate priors** (for a given likelihood) is a key concept in Bayesian analysis and we explore it more fully below.

We are now in a position to inquire about the relative importance of the prior versus the data. Under the assumed prior, the mean (and mode) of the posterior distribution is given by

$$\mu_* = \mu_0 \frac{\sigma_*^2}{\sigma_0^2} + \overline{x} \frac{\sigma_*^2}{\sigma^2/n} \tag{27}$$

Note with a very diffuse prior on μ (i.e., $\sigma_0^2 >> \sigma^2$), that $\sigma_*^2 \to \sigma^2/n$ and $\mu_* \to \overline{x}$. Also note that as we collect enough data, $\sigma_*^2 \to \sigma^2/n$ and again $\mu_* \to \overline{x}$.

Gamma, Inverse-gamma, χ^2, and χ^{-2} Distributions

Before we examine a Gaussian likelihood with unknown variance, a brief aside is needed to develop χ^{-2}, the **inverse chi-square distribution**. We do this via the gamma and inverse-gamma distribution.

The χ^2 is a special case of the **Gamma distribution**, a two parameter distribution. A gamma-distributed variable is denoted by $x \sim \text{Gamma}(\alpha, \beta)$, with density function

$$p(x \,|\, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for } \alpha, \beta, x > 0 \tag{28a}$$

As a function of x, note that

$$p(x) \propto x^{\alpha-1} e^{-\beta x} \tag{28b}$$

We can parameterize a gamma in terms of its mean and variance by noting that

$$\mu_x = \frac{\alpha}{\beta}, \qquad \sigma_x^2 = \frac{\alpha}{\beta^2} \tag{28c}$$

$\Gamma(\alpha)$, the **gamma function** evaluated at α (which normalized the gamma distribution) is defined as

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy \tag{29a}$$

The gamma function is the generalization of the factorial function ($n!$) to all positive numbers, and (as integration by parts will show) satisfies the following identities

$$\Gamma(\alpha) = (1-\alpha)\Gamma(1-\alpha), \qquad \Gamma(1) = 1, \qquad \Gamma(1/2) = \sqrt{\pi} \tag{29b}$$

The χ^2 distribution is a special case of the gamma, with a χ^2 with n degrees of freedom being a gamma random variable with $\alpha = n/2$, $\beta = 1/2$, i.e., $\chi^2_n \sim$ Gamma$(n/2, 1/2)$, giving the density function as

$$p(x \mid n) = \frac{2^{-n/2}}{\Gamma(n/2)} x^{n/2-1} e^{-x/2} \tag{30a}$$

Hence for a χ^2_n,

$$p(x) \propto x^{n/2-1} e^{-x/2} \tag{30b}$$

The **inverse gamma** distribution will prove useful as a conjugate prior for Gaussian likelihoods with unknown variance. It is defined by the distribution of $y = 1/x$ where $x \sim$ Gamma(α, β). The resulting density function, mean, and variance become

$$p(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha-1)} e^{-\beta/x} \quad \text{for } \alpha, \beta, x > 0 \tag{31a}$$

$$\mu_x = \frac{\beta}{\alpha - 1}, \qquad \sigma_x^2 = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \tag{31b}$$

Note for the inverse gamma that

$$p(x) \propto x^{-(\alpha-1)} e^{-\beta/x} \tag{31c}$$

If $x \sim \chi^2_n$, then $y = 1/x$ follows an **inverse chi-square distribution**, and denote this by $y \sim \chi_n^{-2}$. This is a special case of the inverse gamma, with (as for a normal χ^2) $\alpha = n/2$, $\beta = 1/2$. The resulting density function is

$$p(x \mid n) = \frac{2^{-n/2}}{\Gamma(n/2)} x^{-(n/2-1)} e^{-1/(2x)} \tag{32a}$$

with mean and variance

$$\mu_x = \frac{1}{n-2}, \qquad \sigma_x^2 = \frac{2}{(n-2)^2(n-4)} \tag{32b}$$

The **scaled inverse chi-square distribution** is more typically used, where

$$p(x \mid n) \propto x^{-(n/2-1)} e^{-\sigma_0^2/(2x)} \tag{33a}$$

so that the $1/(2x)$ term in the exponential is replaced by an $\sigma_0^2/(2x)$ term. If x follows this distribution, then $\sigma_0^2 \cdot x$ follows a standard χ^{-2} distribution. The scaled-inverse chi-square distribution thus involves two parameters, σ_0^2 and n and it is denoted by SI$-\chi^2(n, \sigma_0^2)$ or $\chi^{-2}_{(n,\sigma_0^2)}$. Note that if

$$x \sim \chi^{-2}_{(n,\sigma_0^2)} \qquad \text{then} \qquad \sigma_0^2\, x \sim \chi_n^{-2} \tag{33b}$$

Table 1. Summary of the functional forms the distributions introduced.

Distribution	$p(x)$/constant
Gamma (α, β)	$x^{\alpha-1}\exp(-\beta x)$
χ^2_n	$x^{n/2-1}\exp(-x/2)$
Inverse-Gamma (α, β)	$x^{-(\alpha-1)}\exp(-\beta/x)$
Inverse-χ^2_n	$x^{-(n/2-1)}\exp[-1/(2x)]$
Scaled Inverse-$\chi^2_{n,S}$	$x^{-(n/2-1)}\exp[-S/(2x)]$

Unknown Variance: Inverse-χ^2 Priors

Now suppose the data are drawn from a normal with known mean μ, but unknown variance σ^2. The resulting likelihood function becomes

$$\ell(\sigma^2 \,|\, \mathbf{x}, \mu) \propto (\sigma^2)^{-n/2} \cdot \exp\left(-\frac{nS^2}{2\sigma^2}\right) \tag{34a}$$

where

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{34b}$$

Notice that since we condition on $\mathbf{x}$ and μ (i.e., their values are known), the S^2 is a constant. Further observe that, as a function of the unknown variance σ^2, the likelihood is proportional to a scaled inverse-χ^2 distribution. Thus, taking the prior for the unknown variance also as a scaled inverse χ^2 with hyperparameters ν_0 and σ_0^2, the posterior becomes

$$\begin{aligned} p(\sigma^2 \,|\, \mathbf{x}, \mu) &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{nS^2}{2\sigma^2}\right)(\sigma^2)^{-\nu_0/2-1} \cdot \exp\left(-\frac{\sigma_0^2}{2\sigma^2}\right) \\ &= (\sigma^2)^{-(n+\nu_0)/2-1} \exp\left(-\frac{nS^2 + \sigma_0^2}{2\sigma^2}\right) \end{aligned} \tag{35a}$$

Comparison to Equation 33a shows that this is also a scaled inverse χ^2 distribution with parameters $\nu_n = (n + \nu_0)$ and $\sigma_n^2 = (nS^2 + \sigma_0^2)$, so that

$$\sigma_n^2\,\sigma^2 \,|\, (\mathbf{x}, \mu) \sim \chi^{-2}_{\nu_n} \tag{35b}$$

Unknown Mean and Variance

Putting all the pieces together, the posterior density for draws from a normal with unknown mean and variance is obtained as follows. First, write the joint prior by conditioning on the variance,

$$p(\mu, \sigma^2) = p(\mu \,|\, \sigma^2) \cdot p(\sigma^2) \tag{36a}$$

As above, assume a scaled inverse chi-square distribution for the variance and, conditioned on the variance, normal prior for the mean with hyperparameters μ_0 and σ^2/κ_0. We write the variance for the conditional mean prior this way because σ^2 is known (as we condition on it) and we scale this by the hyperparameter κ_0. Hence, we assume

$$\sigma^2 \sim \chi^{-2}(\,\nu_0, \sigma_0^2), \qquad (\mu \,|\, \sigma^2) \sim \mathrm{N}\left(\mu_0, \frac{\sigma^2}{\kappa_O}\right) \tag{36b}$$

The resulting posterior marginals become

$$\sigma^2 \,|\, \mathbf{x} \sim \chi^{-2}(\,\nu_n, \sigma_n^2), \quad \text{and} \quad \mu \,|\, \mathbf{x} \sim \mathrm{t}_{\nu_n}\left(\mu_n, \frac{\sigma_n^2}{\kappa_n}\right) \tag{37}$$

where $\mathrm{t}_n(\mu_n, \sigma_n^2)$ denotes a a t-distribution with ν_n degrees of freedom, mean μ_n and variance σ_n^2. Here

$$\nu_n = \nu_0 + n, \qquad \kappa_n = \kappa_0 + n \tag{38a}$$

$$\mu_n = \mu_0 \frac{\kappa_0}{\kappa_n} + \overline{x}\,\frac{n}{\kappa_n} = \mu_0 \frac{\kappa_0}{\kappa_0 + n} + \overline{x}\,\frac{n}{\kappa_0 + n} \tag{38b}$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left(\nu_0\sigma_0^2 + \sum_{i=1}^{n}(\,x_i - \overline{x}\,)^2 + \frac{\kappa_0 n}{\kappa_n}(\,\overline{x} - \mu_0)^2\right) \tag{38c}$$

CONJUGATE PRIORS

The use of a prior density that conjugates the likelihood allows for analytic expressions of the posterior density. Table 2 gives the conjugate priors for several common likelihood functions.

Table 2. Conjugate priors for common likelihood functions.

Likelihood	Conjugate prior
Binomial	Beta
Multinomial	Dirichlet
Poisson	Gamma
Normal	
μ unknown, σ^2 known	Normal
μ known, σ^2 unknown	Inverse Chi-Square
Multivariate Normal	
$\boldsymbol{\mu}$ unknown, $\mathbf{V}$ known	Multivariate Normal
$\boldsymbol{\mu}$ known, $\mathbf{V}$ unknown	Inverse Wishart

We first review some of the additional distributions introduced in Table 2 and then conclude by discussing conjugate priors for members of the exponential family of distributions.

The Beta and Dirichlet Distributions

Where we have frequency data, such as for data drawn from a binomial or muiltinomial likelihood, the **Dirichlet distribution** an appropriate prior. Here, $\mathbf{x} \sim \text{Dirichlet}(\alpha_1, \cdots, \alpha_k)$, with

$$p(x_1, \cdots x_k) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \, x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1} \tag{39a}$$

where

$$\alpha_0 = \sum_{i=1}^{k} \alpha_i, \quad 0 \leq x_i < 1, \quad \sum_{i=1}^{k} x_i = 1, \quad \alpha_i > 0 \tag{39b}$$

where

$$\mu_{x_i} = \frac{\alpha_i}{\alpha_0}, \quad \sigma^2(x_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}, \quad \sigma^2(x_i, x_j) = -\frac{\alpha_i \, \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \tag{39c}$$

An important special case of the Dirichlet is the **Beta distribution**,

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha - 1}(1 - x)^{\beta - 1} \quad \text{for} \quad 0 < x < 1, \quad \alpha, \beta > 0 \tag{40}$$

Wishart and Inverse Wishart Distributions

The **Wishart distribution** can be thought of as the multivariate extension of the χ^2 distribution. In particular, if $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are independent and identically distributed with $\mathbf{x}_i \sim \text{MVN}_k(\mathbf{0}, \mathbf{V})$ – that is, each is drawn from a k-dimensional multivariate normal with mean vector zero and variance-covariance matrix $\mathbf{V}$, then the random ($k \times k$ symmetric, positive definite) matrix

$$\mathbf{W} = \sum_{i=1}^{n} \mathbf{x}_i \, \mathbf{x}_i^T \sim \text{W}_n(\mathbf{V}) \tag{41}$$

Hence, the sum follows a Wishart with n degrees of freedom and parameter $\mathbf{V}$. For the special case of $k = 1$ with $\mathbf{V} = (1)$, this is just a χ_n^2 distribution. The Wishart distribution is the sampling distribution for covariance matrices (just like the χ^2 is associated with the distribution of a sample variance). The probability density function for a Wishart is given by

$$p(\mathbf{W}) = 2^{-nk/2} \pi^{-k(k-1)/k} \, | \, \mathbf{V} \, |^{-n/2} \, | \, \mathbf{W} \, |^{(n+k+1)/2} \, \frac{\exp\left(-\frac{1}{2}\text{tr}\left[\, \mathbf{V}^{-1}\mathbf{W} \,\right]\right)}{\prod_{i=1}^{k} \Gamma\left(\frac{n+1-i}{2}\right)} \tag{42}$$

If $\mathbf{Z} \sim \mathbf{W}_n(\mathbf{V})$, then $\mathbf{Z}^{-1} \sim \mathbf{W}_n^{-1}\left(\mathbf{V}^{-1}\right)$, where $\mathbf{W}^{-1}$ denotes the **Inverse-Wishart distribution**. The density function for an Inverse-Wishart distributed random matrix $\mathbf{W}$ is

$$p(\mathbf{W}) = 2^{-nk/2}\pi^{-k(k-1)/k} \mid \mathbf{V} \mid^{n/2} \mid \mathbf{W} \mid^{-(n+k+1)/2} \frac{\exp\left(-\frac{1}{2}\mathbf{tr}\left[\,\mathbf{V}\mathbf{W}^{-1}\,\right]\right)}{\prod_{i=1}^{k}\Gamma\left(\frac{n+1-i}{2}\right)} \tag{43}$$

Thus, the Inverse-Wishart distribution is the distribution of the inverse of the sample covariance matrix.

Conjugate Priors for the Exponential Family of Distributions

Many common distributions (normal, gamma, poission, binomial,, etc.) are members of the **exponential family**, whose general form is given by Equation 44a. Note that this should not be confused with the simple exponential distribution, which is just one particular member from this family. When the likelihood is in the form of an exponential family, a conjugate prior (also a member of the exponential family of distributions) can be found.

Suppose the likelihood for a single observation (out of n) is in the form of an exponential family,

$$\ell(y_i \mid \theta) = g(\theta)h(y)\exp\left(\sum_{j=1}^{m}\phi_j(\theta)\,t_j(y_i)\right) \tag{44a}$$

Using the prior

$$p(\theta) \propto [\,g(\theta)\,]^b \exp\left(\sum_{j=1}^{m}\phi_j(\theta)\,a_j\right) \tag{44b}$$

yields the posterior density

$$\begin{aligned} p(\theta \mid y) &\propto \left[\prod_{i=1}^{n}\ell(y_i \mid \theta)\right] p(\theta) \\ &=\propto [\,g(\theta)\,]^{b+n}\exp\left(\sum_{j=1}^{m}\phi_j(\theta)\,d_j(y)\right) \end{aligned} \tag{45a}$$

where

$$d_j = a_j + \sum_{i=1}^{n} t_j(y_i) \tag{45b}$$

Thus Equation 44b is the conjugate prior density for the likelihood given by Equation 44a, with the posterior having the same form as the prior, with $n + b$ (in the posterior) replacing b and d_j replacing a_j.

References

Draper, David. 2000. *Bayesian Hierarchical Modeling*. Draft version can be found on the web at **http://www.bath.ac.uk/∼masdd/**

Efron, B. 1986. Why isn't everyone a bayesian? *American Statistician* 40: 1-11.

Glymour, C. 1981. Why I am not a Bayesian, in *The philosophy of science*, *ed.* by D. Papineau. Oxford University Press.

Jeffreys, H. S. 1961. *Theory of Probability*, 3rd ed. Oxford University Press.

Lee, P. M. 1997. *Bayesian statistics: An introduction*, 2nd ed. Arnold, London.

Lindley, D. V. 1965. *Introduction to Probability and Statistics from a Bayesian Viewpoint* (2 Volumes), University Press, Cambridge.

Stigler, S. M. 1983. Who discovered Bayes's theorem? *American Statistician* 37: 290–296

Tanner, M. A. 1996. *Tools for statistical inference: Methods for exploration of posterior distributions and likelihood functions*, 3rd ed. Springer-Verlag.