Appendix 2

Introduction to Bayesian Analysis

A form of inference which regards parameters as being random variables possessed of prior distributions reflecting the accumulated state of knowledge — Kendall and Buckland (1971)

Draft version 18 April 2014

The history of statistical methods in genetics closely parallels advances in computation. Before the widespread use of computers, method-of-moments approaches were common as they are relatively easy to compute. Here, a summary statistic of the data is computed whose expected value is the parameter of interest. In the mid-1970's, maximum-likelihood (ML) methods become much more common place, as they offer a very flexible platform for statistical analysis (estimation, determining precision, hypothesis testing), but at the cost of numerically searching an often highly complex multidimensional likelihood surface (LW Appendix 4). Both these approaches typically return point estimators for variables of interest, along with some measure of their uncertainty. As opposed these classical (or **frequentist**) methods, **Bayesian statistics** (which can be viewed as a natural extension of likelihood methods) is concerned with generating the full *distribution* for the parameters $\boldsymbol{\Theta}$ given the data $\mathbf{x}$, i.e., the posterior distribution $p(\boldsymbol{\Theta} \mid \mathbf{x})$. 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 any number of introductory texts (e.g, Lindley 1965; Berger 1985; Carlin and Louis 2000; Lee 2012; Gelman et al. 2013) for a more complete introduction, and to Sorensen and Gianola (2002) for applications to classical quantitative genetics. While very deep (and very subtle) differences in philosophy separate hard-core **Bayesians** from hard-core **frequentists** (Efron 1986; Glymour 1981), our treatment of Bayesian methods is motivated simply by their use as a powerful statistical tool. This appendix focuses on the basic theory, while the computational approaches that make these methods feasible are examined in Appendix 3.

WHY ARE BAYESIAN METHODS BECOMING MORE POPULAR?

In addition to providing a more formal framework for dealing with parameter uncertainty, two specific features have fueled the rapid growth of Bayesian approaches in genetics and genomics. First, under a Bayesian analysis, all parameters are random (as opposed to fixed) effects (Chapter 19). This has profound implications for degrees of freedom. Consider a microarray experiment with 30,000 features (genes of interest) whose expression levels are contrasted over a set of 100 normal versus 100 cancerous liver cells. Treating the differential expression level of any particular gene as a fixed effect (an unknown constant to be estimated) very quickly uses all of the degrees of freedom given the small sample size. Conversely, if these are treated as **random effects**, with the expression difference associated with a particular gene being a random variable drawn from some underlying (and unknown) distribution, then the only degrees of freedom lost are those used to estimate the associated parameters for this underlying distribution. Further, prediction of the random realization that corresponds to a particular gene borrows information over all genes. Thus, a Bayesian analysis can handle high-dimensional experiments where the number of parameters p greatly exceeds the

number of observations n, in a framework that fully manages the uncertainly over all these estimates. Second, they are computationally feasible, as approaches such MCMC (Appendix 3) allow high-dimensional datasets to be analyzed in a computationally-efficient manner. In settings with a large number of nuisance parameters, or a high-dimensional dataset, a Bayesian approach not only has considerable appeal, it may be the only approach that is feasible.

BAYES' THEOREM

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

$$\Pr(\theta \mid x) = \frac{\Pr(x, \theta)}{\Pr(x)} \tag{A2.1a}$$

where (for now) x and θ are discrete random variables. Again from the definition of conditional probability, we can express the joint probability by conditioning on θ to give

$$\Pr(x, \theta) = \Pr(x \mid \theta) \Pr(\theta) \tag{A2.1b}$$

Putting these together gives Bayes' theorem:

$$\Pr(\theta \mid x) = \frac{\Pr(x \mid \theta) \Pr(\theta)}{\Pr(x)} \tag{A2.2a}$$

Bayes' theorem flips the conditioning variable, allowing us to move from $\Pr(x \mid \theta)$ to $\Pr(\theta \mid x)$. With n possible outcomes $(\theta_1, \cdots, \theta_n)$,

$$\Pr(\theta_j \mid x) = \frac{\Pr(x \mid \theta_j) \Pr(\theta_j)}{\Pr(x)} = \frac{\Pr(x \mid \theta_j) \Pr(\theta_j)}{\sum_{i=1}^{n} \Pr(\theta_i)\ \Pr(x \mid \theta_i)} \tag{A2.2b}$$

In Bayesian statistics, x represents an observable variable (the data), while θ represents a parameter describing the distribution of x. In this setting $\Pr(\theta)$ is the **prior distribution** of possible parameter values, while $\Pr(\theta \mid x)$ is the subsequent **posterior distribution** of θ given the observed data x. In classical statistics, the unknown parameters are treated as fixed and the data are considered random, while under a Bayesian analysis, the data are considered fixed and the unknown parameters that generated them as random.

All of the above statements hold for continuous random variables, with the probability density function p replacing the discrete probability $\Pr$. In particular, the continuous multivariate version of Bayes' theorem is

$$p(\boldsymbol{\Theta} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{p(\mathbf{x})} = \frac{p(\mathbf{x} \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{\int p(\mathbf{x}, \boldsymbol{\Theta})\, d\boldsymbol{\Theta}} \tag{A2.3}$$

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 $p(\boldsymbol{\Theta} \mid \mathbf{x})$ is the posterior distribution given the prior $p(\boldsymbol{\Theta})$ and the data $\mathbf{x}$.

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). This is an example of **Stigler's Law of Eponymy** (Stigler 1980), wherein no discovery or invention is named after its first discoverer (an **eponym**). As is fitting, Stigler's law is self-consistent, as this phenomena was first noticed by Merton (1965).

Example A2.1. Suppose one in every 1000 families has a genetic disorder (sex-bias) in which they produce only female offspring. For any particular family, 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 this 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 $x =$ five girls, and

$$\text{Pr(five girls} \mid \text{sex bias family)} = 1, \quad \text{Pr(five girls} \mid \text{normal family)} = (1/2)^5$$

Hence, $\Pr(x = 5 \mid \theta = 1)$ = 1, while $\Pr(x = 5 \mid \theta = 0)$ = $(1/2)^5$. 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),

$$\text{Pr(5 girls) = Pr(5 girls} \mid \text{normal)*Pr(normal) + Pr(5 girls} \mid \text{sex-bias)*Pr(sex-bias)}$$

giving

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

Hence,

$$\Pr(\theta = 1 \mid x = 5 \text{ girls}) = \frac{\Pr(x \mid \theta = 1)\Pr(\theta = 1)}{\Pr(x)} = \frac{1 \cdot 0.001}{0.0322} = 0.031,$$

showing that a family of five with all girls is 31 times more likely than a random family to have the sex-bias disorder.

Example A2.2. 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 $\varphi(x \mid \mu, 1) = (2\pi)^{-1/2} e^{-(x-\mu)^2/2}$ denote the density function for a normal distribution 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)	p(x|G)	Pr(G)·p(x|G)
QQ	0.3	$\varphi(3 \mid 2.1, 1) = 0.266$	0.078
Qq	0.2	$\varphi(3 \mid 3.5, 1) = 0.350$	0.070
qq	0.5	$\varphi(3 \mid 1.3, 1) = 0.094$	0.047

Since $p(x) = \sum_G$ Pr(G)·p(x | G) = 0.195, Bayes' theorem gives the posterior probabilities for the genotypes given the observed value of 3 as:

$$\Pr(QQ \mid x = 3) = 0.078/0.195 = 0.409$$

$$\Pr(Qq \mid x = 3) = 0.070/0.195 = 0.361$$

$$\Pr(qq \mid x = 3) = 0.047/0.195 = 0.241$$

Thus, there is a 41 percent chance this individual has genotype QQ, a 36 percent chance it is Qq, and only a 24 percent chance it is qq.

FROM LIKELIHOOD TO BAYESIAN ANALYSIS

The method of maximum likelihood (LW Appendix 4) 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 of the likelihood function (the maximal value of ℓ, as a function of $\boldsymbol{\Theta}$ given the data $\mathbf{x}$), and use the local curvature around the mode 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 estimators and that most LR tests follow χ^2 distributions. These nice features don't necessarily hold for small samples. Coversely, a Bayesian analysis is *exact* for any sample size given a specificed prior.

To transition from a likelihood to a Bayesian analysis, we start with some prior distribution $p(\boldsymbol{\Theta})$ capturing our initial knowledge/best guess about the possible values of the unknown parameter(s). From Bayes' theorem, the data (likelihood) is combined with the prior to produce a posterior distribution,

$$p(\boldsymbol{\Theta} \mid \mathbf{x}) = \frac{1}{p(\mathbf{x})} \cdot p(\mathbf{x} \mid \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}) \tag{A2.4a}$$

$$= (\text{normalizing constant}) \cdot p(\mathbf{x} \mid \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta}) \tag{A2.4b}$$

$$= \text{constant} \cdot \text{likelihood} \cdot \text{prior} \tag{A2.4c}$$

as $p(\mathbf{x} \mid \boldsymbol{\Theta}) = \ell(\boldsymbol{\Theta} \mid \mathbf{x})$ is just the likelihood function (LW Appendix 4) and $1/p(\mathbf{x})$ is a constant (with respect to $\boldsymbol{\Theta}$). Because of this, the posterior distribution is often written as

$$p(\boldsymbol{\Theta} \mid \mathbf{x}) \propto \ell(\boldsymbol{\Theta} \mid \mathbf{x})\, p(\boldsymbol{\Theta}) \tag{A2.4d}$$

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,

$$p(\mathbf{x}) = \int_{\boldsymbol{\Theta}} p(\mathbf{x} \mid \boldsymbol{\Theta}) \cdot p(\boldsymbol{\Theta})\, d\boldsymbol{\Theta} \tag{A2.5}$$

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 (the curvature of the likelihood surface). If the posterior is highly dependent on the prior, then the data likely has little signal (a **flat likelihood surface**), while if

the posterior is largely unaffected by different priors, the data are likely highly informative (a sharply peaked likelihood surface). To see this, taking logs on Equation A2.4c (and ignoring the normalizing constant) gives

$$\log(\text{posterior}) = \log(\text{likelihood}) + \log(\text{prior}) \tag{A2.6}$$

When the likelihood signal is strong, it largely dominates the prior in the resulting posterior, but when a likelihood is weak, the prior can dominate.

Marginal Posterior Distributions

Often only a subset of the unknown parameters is really of concern to us, the rest being **nuisance parameters** that are of no interest, but still must be fitted in the model. A very strong feature of Bayesian analysis is that we can account for all the uncertainty introduced into parameters of interest by uncertainty in the values of nuisance parameters. This is accomplished by integrating the nuisance parameters 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, which is not fully captured by standard classical approaches. 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 resulting marginal posterior for σ^2 captures all of the uncertainty in the estimation of μ that influences the uncertainty in σ^2. This is an especially nice feature when a large number of nuisance parameters must be estimated.

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

$$p(\boldsymbol{\Theta}_1 \,|\, \mathbf{y}) = \int_{\boldsymbol{\Theta}_{nu}} p(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_{nu} \,|\, \mathbf{y})\, d\boldsymbol{\Theta}_{nu} \tag{A2.7}$$

While these complex integrals appear quite daunting (and indeed almost always are from an analytic standpoint), generating draws from the marginal distribution is usually very straightforward using the MCMC methods examined in Appendix 3.

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 posterior distribution** (its maximal value),

$$\widehat{\theta} = \max_{\theta} [\, p(\,\theta \,|\, \mathbf{x}\,)] \tag{A2.8a}$$

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

$$\widehat{\theta} = E[\,\theta \,|\, \mathbf{x}\,] = \int \theta\, p(\,\theta \,|\, \mathbf{x}\,) d\theta \tag{A2.8b}$$

Another candidate is the **median of the posterior**, which is more robust than the mean to outliers. Here the estimator satisfies $\Pr(\theta > \widehat{\theta} \,|\, \mathbf{x}) = \Pr(\theta < \widehat{\theta} \,|\, \mathbf{x}) = 0.5$, hence

$$\int_{\widehat{\theta}}^{+\infty} p(\,\theta \,|\, \mathbf{x}\,) d\theta = \int_{-\infty}^{\widehat{\theta}} p(\,\theta \,|\, \mathbf{x}\,) d\theta = \frac{1}{2} \tag{A2.8c}$$

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 in Appendix 3, we can generally obtain the posterior by simulation using MCMC sampling, and hence the Bayesian estimate of a parameter is frequently presented as a frequency histogram (potentially smoothed) of the MCMC-generated samples from 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

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

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 analysis. 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, $(1-\alpha)$ of the time the confidence intervals will enclose the unknown parameter. Thus, it is a measure of the *frequency* of times in independent experiments that the CI encloses the true value (and hence the term frequentist for this type of statistics). In contrast, with a Bayesian HDR, there is a $(1-\alpha)$ probability that the interval contains the true value of the unknown parameter. While these two intervals at first blush appear to be essentially identical, they are not and indeed are fundamentally (but subtly) different. Often the CI and Bayesian intervals contain essentially the same values, 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 typically avoid using the term confidence interval, using credible interval instead.

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 significance of a particular hypothesis, and (if possible) compute a p-value (the

probability p of observing a value equal to, or more extreme than, that of the test statistic if the null hypothesis is indeed correct). At first blush, one would think that the idea of a hypothesis test is trivial in a Bayesian framework, as using the posterior distribution gives us expected p values, e.g.,

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

The fault in this logic under a Bayesian framework 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 suggested by our prior. For example, suppose that 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 influence hypothesis testing*. To formalize this idea, let

$$p_0 = \Pr(\theta \in \Theta_0 \,|\, \mathbf{x}), \qquad p_1 = \Pr(\theta \in \Theta_1 \,|\, \mathbf{x}) \tag{A2.9a}$$

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

$$\pi_0 = \Pr(\theta \in \Theta_0) \qquad \text{and} \qquad \pi_1 = \Pr(\theta \in \Theta_1) \tag{A2.9b}$$

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,

$$B_0 = \frac{p_0/p_1}{\pi_0/\pi_1} = \frac{p_0 \pi_1}{p_1 \pi_0} \tag{A2.10a}$$

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

$$B_0 = \frac{p_0(1 - \pi_0)}{\pi_0(1 - p_0)} \tag{A2.10b}$$

By symmetry note that the Bayes factor B_1 in favor of H_1 versus H_0 is just $B_1 = 1/B_0$.

Example A2.3. Consider our first example from above where the prior and posterior probabilities for the null were $\pi_0 = 0.1$ and $p_0 = 0.05$ (respectively). The Bayes factor in favor of H_1 versus H_0 is

$$B_1 = \frac{\pi_0(1 - p_0)}{p_0(1 - \pi_0)} = \frac{0.1 \cdot 0.95}{0.05 \cdot 0.9} = 4.22$$

Similarly, for the second example where the prior for the null was $\pi_0 = 0.5$,

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

Here, the data gave close to a twenty-fold improvement (relative to the prior) in support of H_1. Bayes factors and p values represent fundamentally different approaches to an analysis and are not formally comparable. However, a loose interpretation is that a factor of 20 is akin to the level of support of a $p = 0.05$ and a factor of 100 to $p = 0.01$.

When the hypotheses are simple (i.e., single values), 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

$$\frac{p_0}{p_1} = \frac{\pi_0\, p(\mathbf{x} \,|\, \theta_0)}{\pi_1\, p(\mathbf{x} \,|\, \theta_1)} \tag{A2.11a}$$

and from Equation A2.10a the Bayes factor (in favor of the null) reduces the

$$B_0 = \frac{p(\mathbf{x} \,|\, \theta_0)}{p(\mathbf{x} \,|\, \theta_1)} \tag{A2.11b}$$

which is simply a *likelihood ratio* (LW Appendix 4).

When 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

$$p_i(\theta) = p(\theta)/\pi_i \qquad \text{for} \quad i = 0, 1 \tag{A2.12}$$

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

$$\begin{aligned} p_i = \Pr(\theta \in \Theta_i \,|\, \mathbf{x}) &= \int_{\theta \in \Theta_i} p(\theta \,|\, \mathbf{x}) d\theta \\ &\propto \int_{\theta \in \Theta_i} p(\theta) p(\mathbf{x} \,|\, \theta) d\theta \\ &= \pi_i \int_{\theta \in \Theta_i} p(\mathbf{x} \,|\, \theta) p_i(\theta) d\theta \end{aligned} \tag{A2.13}$$

where the second step follows from Bayes' theorem and the final step follows from Equation A2.12. The Bayes factor in favor of the null hypothesis becomes

$$B_0 = \left(\frac{p_0}{\pi_0}\right)\left(\frac{\pi_1}{p_1}\right) = \frac{\int_{\theta \in \Theta_0} p(\mathbf{x} \,|\, \theta) p_0(\theta) d\theta}{\int_{\theta \in \Theta_1} p(\mathbf{x} \,|\, \theta) p_1(\theta) d\theta}, \tag{A2.14}$$

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

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 a sufficient signal, even a bad prior will still not greatly influence the posterior. In a sense, this is an asymptotic (large-sample) property of Bayesian analysis in that all but pathological priors can be overcome by sufficient amounts of data. As mentioned above, one can check the impact of the prior by assessing the stablity of posterior over a collection of diverse priors. 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 posterior, 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 shortly, but first discuss other approaches for construction of priors.

Diffuse Priors

One of the most common priors is the **flat** or **diffuse** (also called **uninformative** or **naive**) prior, which is simply a constant,

$$p(\theta) = k = \frac{1}{b-a} \qquad \text{for} \qquad a \leq \theta \leq b \tag{A2.15a}$$

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,

$$p(\theta \,|\, \mathbf{x}) = C\, \ell(\theta \,|\, \mathbf{x}) \tag{A2.15b}$$

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 (i.e., parameter) 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 (i.e., assuming $p[\theta \,|\, \mathbf{x}] \propto \ell[\theta \,|\, \mathbf{x}]$) is referred to as an **improper prior**, and care must be taken to ensure that the product of the prior and the likelihoood results in a proper posterior (i.e., it has a finite integral over the parameter range).

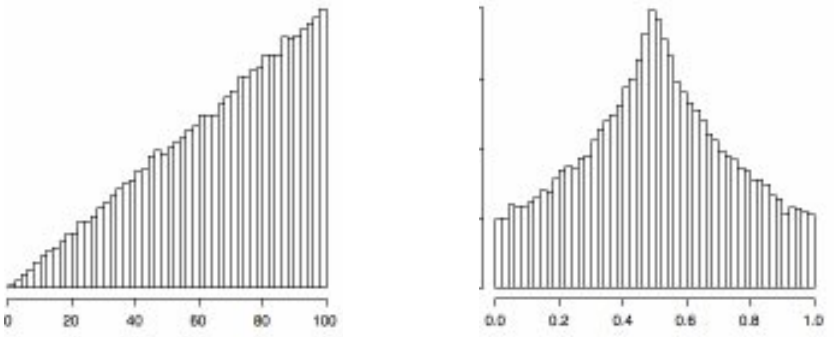

Figure A2.1. A uniform prior on one scale does not result in a flat prior on a transformed scale. Suppose a flat prior on (0,10000) is assumed for both the additive and residual variances. **Left:** The resulting prior (over 0,100) for the standard deviation of either variance. **Right:** The resulting prior for h^2.

Another problem with a uniform prior is when the questions of interest reside on a different scale. A variable uniform on one scale may be far from uniform on a transformed scaled. Figure A2.1 shows two examples based on the assumption of a flat prior on the variance. A uniform prior on the variance does *not* result in a uniform prior on the standard deviation (e.g., Van Dongen 2006). Likewise if one assumes that the additive and residual variances have flat priors, this does not imply a flat prior for h^2, but rather a prior that

is sharply peaked at 1/2. When assuming a flat prior, care most be taken that it is truly uninformative on the appropriate scale of biological interest. Otherwise the choice of what superficially appears as an unbiased prior may instead create a bias that the signal in the data has to overcome.

The Jeffreys' Prior

Jeffreys (1961) proposed a general prior based on the Fisher information I of the likelihood. Recall (LW Appendix 4) that

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

The **Jeffreys' prior** is given

$$p(\theta) \propto \sqrt{I(\theta \,|\, \mathbf{x}\,)} \tag{A2.16}$$

A full discussion, with derivation, can be found in Lee (2012).

Example A2.4. Consider the likelihood of x successes in 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

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

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

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

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 (Equation A2.38).

When there are k parameters, $\mathbf{I}$ is the $k \times k$ Fisher Information 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

$$p(\boldsymbol{\Theta}) \propto \sqrt{\det[\mathbf{I}(\theta \,|\, \mathbf{x})]} \tag{A2.17}$$

Example A2.5. Suppose our data consists of n independent draws from a normal distribution with unknown mean and variance, μ and σ^2. In LW Appendix 4, 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 ($p(\mu) = \text{constant}$). Note here 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, as well as treating a common assumption in quantitative genetics, consider the case where data are drawn from a normal (Gaussian) distribution, so that the likelihood function for the ith observation x_i is

$$\ell(\mu, \sigma^2 \,|\, x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{A2.18a}$$

Assuming independence, the resulting full likelihood for all n data points is

$$\ell(\mu \,|\, \mathbf{x}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right) \tag{A2.18b}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu n\overline{x} + n\mu^2\right)\right] \tag{A2.18c}$$

Gaussian Likelihood with 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

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \tag{A2.19}$$

The mean and variance of the prior, μ_0 and σ_0^2 are referred to as **hyperparameters**. Here, μ_0 specifies a prior location for the parameter, while σ_0^2 specifies our uncertainty in this prior location – the larger σ_0^2, the greater our uncertainty. In the limit as $\sigma_0^2 \to \infty$, $p(\mu)$ approaches a flat (and in this case, improper) prior.

A useful device 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

$$p(\boldsymbol{\Theta}_2 \,|\, \mathbf{x}, \boldsymbol{\Theta}_1) = f(\mathbf{x}, \boldsymbol{\Theta}_1) \cdot g(\mathbf{x}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \tag{A2.20a}$$

then

$$p(\boldsymbol{\Theta}_2 \,|\, \mathbf{x}, \boldsymbol{\Theta}_1) \propto g(\mathbf{x}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \tag{A2.20b}$$

which follows since $f(\mathbf{x}, \boldsymbol{\Theta}_1)$ is constant with respect to $\boldsymbol{\Theta}_2$.

With the prior given by Equation A2.19, we can express the resulting posterior distribution as

$$\begin{aligned} p(\mu \,|\, \mathbf{x}) &\propto \ell(\mu \,|\, \mathbf{x}) \cdot p(\mu) \\ &\propto \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2} - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} x_i^2 - 2\mu n\overline{x} + n\mu^2 \right] \right) \end{aligned} \tag{A2.21a}$$

We can factor out additional terms not involving μ to obtain

$$p(\mu \,|\, \mathbf{x}) \propto \exp\left(-\frac{\mu^2}{2\sigma_0^2} + \frac{\mu\,\mu_0}{\sigma_0^2} + \frac{\mu n\overline{x}}{\sigma^2} - \frac{n\mu^2}{2\sigma^2} \right) \tag{A2.21b}$$

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

$$-\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) + \mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2} \right) = -\frac{\mu^2}{\sigma_*^2} + \frac{2\mu\mu_*}{2\sigma_*^2} \tag{A2.22a}$$

where

$$\sigma_*^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \quad \text{and} \quad \mu_* = \sigma_*^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\overline{x}}{\sigma^2} \right) \tag{A2.22b}$$

Finally, by completing the square, we have

$$p(\mu \,|\, \mathbf{x}) \propto \exp\left(-\frac{(\mu - \mu_*)^2}{2\sigma_*^2} + f(\mathbf{x}, \mu_0, \sigma^2, \sigma_0^2) \right) \tag{A2.22c}$$

Recalling Equation A2.20b, we can ignore the second term in the exponential (as it does not involve μ) and the resulting posterior for μ becomes

$$p(\mu \,|\, \mathbf{x}) \propto \exp\left(-\frac{(\mu - \mu_*)^2}{2\sigma_*^2} \right) \tag{A2.23a}$$

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

$$\mu \,|\, (\mathbf{x}, \sigma^2) \sim \mathrm{N}\left(\mu_*, \sigma_*^2 \right) \tag{A2.23b}$$

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 (but with different distribution parameters). The use of such **conjugate priors** associated with a given family of likelihood functions 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

$$\mu_* = \mu_0 \frac{\sigma_*^2}{\sigma_0^2} + \overline{x} \frac{\sigma_*^2}{\sigma^2/n} \tag{A2.24}$$

With a very diffuse prior on μ (i.e., $\sigma_0^2 \gg \sigma^2$), $\sigma_*^2 \to \sigma^2/n$ and $\mu_* \to \overline{x}$. Also note from Equation A2.22b that as we collect enough data (i.e., sufficiently large n), $\sigma_*^2 \to \sigma^2/n$ and again $\mu_* \to \overline{x}$.

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

Before examining the 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 distributions, as both χ^2 and χ^{-2} are special cases of these.

To motivate the gamma distribution, first consider the simple exponential waiting-time distribution, where if β the **rate** (the probability of a success in some small time unit δ_t is given by $\beta\,\delta_t$), then the pdf for the exponential is

$$p(x \mid \beta) = \beta e^{-\beta x} \quad \text{for} \quad 0 \le x < \infty, \quad \beta > 0$$

Since the expected waiting time until a success $\lambda = 1/\beta$, this can be reparameterized in terms of the **scale** (waiting time) parameter as

$$p(x \mid \beta) = \lambda^{-1} e^{-x/\lambda}$$

The sum of k exponentials with the same rate (or scale) parameter is called an **Erlang distribution**, which was initially developed for certain problems in telephone queueing theory. Expressed in terms of the rate parameter,

$$p(x \mid k, \beta) = \frac{\beta^k}{(k-1)!}\, x^{k-1}\, e^{-\beta x} \quad \text{for} \quad 0 \le x < \infty$$

where an integer k is the **shape parameter**, with $k = 1$ recovering the exponental. The gamma distribution follows by allowing the shape parameter to be any positive number α, with $x \sim \text{Gamma}(\alpha, \beta)$ having density function

$$p(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-\beta x} \quad \text{for } \alpha, \beta, x > 0 \tag{A2.25a}$$

where the factorial in the Erlang is replaced by the gamma function $\Gamma(x)$, defined below (Equation A2.26a). As a function of x, note that

$$p(x \mid \alpha, \beta) \propto x^{\alpha-1} e^{-\beta x} \tag{A2.25b}$$

Expressed in terms of the scale ($\lambda = 1/\beta$) parameter, the pdf becomes

$$p(x \mid \alpha, \lambda) = \frac{\lambda^{-\alpha}}{\Gamma(\alpha)}\, x^{\alpha-1} e^{-x/\lambda}$$

giving

$$p(x \mid \alpha, \lambda) \propto x^{\alpha-1} e^{-x/\lambda} \tag{A2.25b}$$

Since both the rate and scale versions of the gamma distribution are used widely used, take care to know which version your software package is using (R uses the scale parameter version). We can parameterize a gamma in terms of its mean and variance by noting that

$$\mu_x = \frac{\alpha}{\beta} = \alpha\ \lambda \qquad \sigma_x^2 = \frac{\alpha}{\beta^2} = \alpha\,\lambda^2 \tag{A2.25c}$$

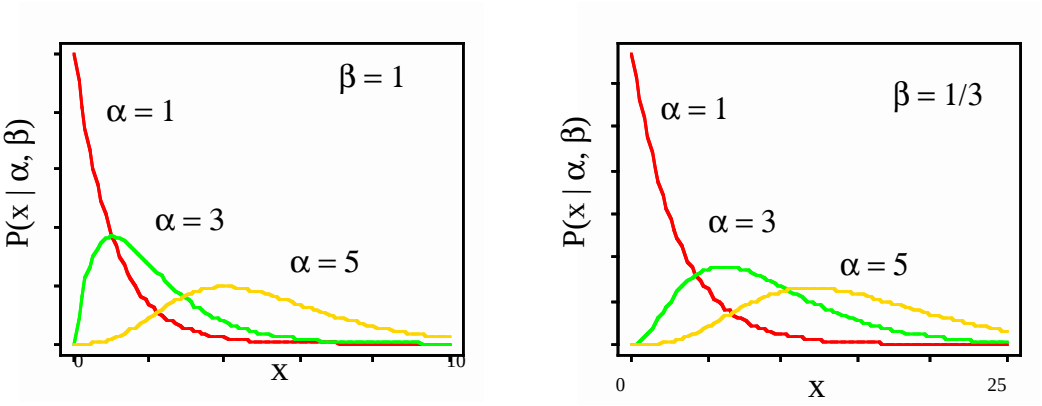

Figure A2.2. The effect of the shape (α) and rate ($\beta = 1/\lambda$, the inverse of the scale) parameters on the gamma. For $\alpha = 1$, it is the simple monotonically decreasing exponential, while for $\alpha > 1$ it is unimodal. The affect of a change in the rate/scale is to keep the general shape, but change the scaling with respect to x.

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

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy \tag{A2.26a}$$

This is the generalization of the factorial function from integers to all positive numbers. If n is an integer, then $\Gamma(n) = (n-1)!$ Using integration by parts, one can show that Γ satisfies the following identities

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha), \qquad \Gamma(1) = 1, \qquad \Gamma(1/2) = \sqrt{\pi} \tag{A2.26b}$$

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

$$p(x \mid n) = \frac{2^{-n/2}}{\Gamma(n/2)}\, x^{n/2-1} e^{-x/2} \tag{A2.27a}$$

Hence for $x \sim \chi^2_n$,

$$p(x) \propto x^{n/2-1} e^{-x/2} \tag{A2.27b}$$

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 $x \sim \text{Gamma}(\alpha, \beta)$. The resulting density function, mean, and variance become

$$p(x \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, x^{-(\alpha+1)} e^{-\beta/x} \quad \text{for } \alpha, \beta, x > 0 \tag{A2.28a}$$

The mean and variance for this distribution are only defined (finite) if α is sufficiently large,

$$\mu_x = \frac{\beta}{\alpha - 1}, \text{ for } \alpha > 1; \qquad \sigma_x^2 = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \text{ for } \alpha > 2 \tag{A2.28b}$$

Note for the inverse gamma that

$$p(x \mid \alpha, \beta) \propto x^{-(\alpha+1)} e^{-\beta/x} \tag{A2.28c}$$

If $y \sim \chi_n^2$, then $x = 1/y$ follows an **inverse chi-square distribution**, and denote this by $x \sim \chi_n^{-2}$. 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

$$p(x \mid n) = \frac{2^{-n/2}}{\Gamma(n/2)} x^{-(n/2+1)} e^{-1/(2x)} \tag{A2.29a}$$

with mean and variance

$$\mu_x = \frac{1}{n-2}, \qquad \sigma_x^2 = \frac{2}{(n-2)^2(n-4)} \tag{A2.29b}$$

The **scaled inverse chi-square distribution** is more typically used, where the rate parameter β (which equals $1/2$ under a chi-square) is replaced $\beta = \sigma_0^2/2$,

$$p(x \mid n) \propto x^{-(n/2+1)} e^{-\sigma_0^2/(2x)} \tag{A2.30a}$$

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

$$x \sim \chi^{-2}_{(n,\sigma_0^2)} \qquad \text{then} \qquad \sigma_0^2 \cdot x \sim \chi_n^{-2}, \tag{A2.30b}$$

showing that σ_0^2 is a scaling factor on a standard ($\beta = 1/2$) inverse chi-square.

Table A2.1. Summary of the functional forms in terms of x of various gamma-related distributions.

Distribution	α	β	$p(x)$/constant
Gamma (α, β)			$x^{\alpha-1}\exp(-\beta x)$
χ_n^2	$n/2$	$1/2$	$x^{n/2-1}\exp(-x/2)$
Inverse-Gamma (α, β)			$x^{-(\alpha+1)}\exp(-\beta/x)$
Inverse-χ_n^2	$n/2$	$1/2$	$x^{-(n/2+1)}\exp[-1/(2x)]$
Scaled Inverse-χ^2_{n,σ_0^2}	$n/2$	$\sigma_0^2/2$	$x^{-(n/2+1)}\exp[-\sigma_0^2/(2x)]$

Gaussian Likelihood With Unknown Variance: Scaled Inverse-χ^2 Priors

Suppose data are drawn from a normal with known mean μ, but unknown variance σ^2. The resulting likelihood function can be expressed as

$$\ell(\sigma^2 \mid \mathbf{x}, \mu) \propto (\sigma^2)^{-n/2} \cdot \exp\left(-\frac{nS^2}{2\sigma^2}\right) \tag{A2.31a}$$

where

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{A2.31b}$$

Notice that since we condition on $\mathbf{x}$ and μ (i.e., their values are known), 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 (Equation A2.30a). 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

$$\begin{aligned} p(\sigma^2 \,|\, \mathbf{x}, \mu) &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{nS^2}{2\sigma^2}\right)(\sigma^2)^{-\nu_0/2-1} \cdot \exp\left(-\frac{\sigma_0^2}{2\sigma^2}\right) \\ &= (\sigma^2)^{-(n+\nu_0)/2-1} \exp\left(-\frac{nS^2 + \sigma_0^2}{2\sigma^2}\right) \end{aligned} \tag{A2.32a}$$

Comparison to Equation A2.30a shows the resulting posterior is also a scaled inverse χ^2 distribution with parameters $\nu_n = (n + \nu_0)$ and $\sigma_n^2 = (nS^2 + \sigma_0^2)$. Hence,

$$\text{for the prior } \sigma^2 \sim \chi^{-2}_{\nu_0,\sigma_0^2}, \quad \sigma_n^2 \cdot \sigma^2 \,|\, (\mathbf{x}, \mu) \sim \chi^{-2}_{\nu_n} \tag{A2.32b}$$

Student's t Distribution

The final distribution needed for a Bayesian analysis of a Gaussian likelihood is the $\mathbf{t}$ (or **Student's t**) distribution. Suppose that $x_i \sim N(\mu, \sigma^2)$, so that for n independent draws, $\overline{x} \sim N(\mu, \sigma^2/n)$, implying $(\overline{x} - \mu)/\sqrt{\sigma^2/n} \sim U$, where $U \sim N(0, 1)$ denotes a unit normal. Likewise, the sample variance $\text{Var}(x)$ follows a scaled chi-square distribution, with $\text{Var}(x) \sim (n-1)\sigma^2\chi^2_{n-1}$ (LW Equation A5.14c). When the estimated variance $\text{Var}(x)$ is used in place of the true variance σ^2, $(\overline{x} - \mu)/\sqrt{\text{Var}(x)/n}$ now follows a t distribution with $n-1$ degrees of freedom, giving raise to the very familiar *t-test*. Notice that

$$t_{n-1} = \left(\frac{\overline{x} - \mu}{\sigma/\sqrt{n}}\right)\left(\frac{1}{\sqrt{\text{Var}/\sigma^2}}\right) = \frac{U}{\sqrt{\chi^2_{n-1}/(n-1)}}$$

Thus, a t_ν random variable follows the distribution of a unit normal divided by the square root of a scaled chi-square with ν degrees of freedom,

$$t_\nu = \frac{U}{\sqrt{\chi^2_\nu/\nu}} \tag{A2.33a}$$

Note that $E(\chi^2_\nu) = \nu$, so that $E(\chi^2_\nu/\nu) = 1$. Relative to a normal, a t distribution is more peaked with heavier tails, and this kurtosis becomes more pronounced as ν decreases. Indeed, the tails fall off sufficiently slowly that a t with two degrees of freedom has an infinite variance, while a t with four (or fewer) degrees of freedom has an infinite fourth moment. The coefficient of kurtosis (LW Equation 2.12a) for a t with $\nu > 4$ degrees of freedom is $k_4 = 6/(\nu - 4)$, which approaches the value (zero) for a normal for large ν. For $\nu > 30$, the t essentially becomes a unit normal distribution.

As with a unit normal, one can also add scale and location to a standard t_ν, generating a three-parameter family of distributions,

$$t_\nu(\mu, \sigma) = \mu + \sigma \cdot t_\nu \tag{A2.33b}$$

The resulting mean and variance are

$$E[t_\nu(\mu, \sigma)] = \mu, \quad \sigma^2[t_\nu(\mu, \sigma)] = \sigma^2 \frac{\nu}{\nu - 2} \quad \text{for } \nu > 2 \tag{A2.33c}$$

Hence, μ and σ control the location and scale (uncertainty about the location), while ν controls the kurtosis, with heavy tails for ν small and little/no kurtosis for $\nu > 20$. The resulting density function becomes

$$p(x \mid \nu, \mu, \sigma) = \frac{\Gamma[(\nu+1)/2]}{\Gamma[\nu/2]\sigma\sqrt{\pi\,\nu}} \left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-(\nu+1)/2} \tag{A2.33d}$$

The role of the t distribution in Bayesian statistics is two-fold. First, it is often used as a *more robust prior*, as its heavier tails may better account for outliers. Using a t distribution with low degrees of freedom (often $\nu = 5$) offers a prior that is similar to a normal, but allows for more frequent extreme values. The second scenario is that the marginal posterior for μ of a Gaussian likelihood with a normal prior on the mean and an inverse chi-square on the variance is a t distribution. This arises after the joint posterior is integrated over all possible σ^2 values (i.e., over an inverse chi-square).

General Gaussian Likelihood: Unknown Mean and Variance

Putting all these 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,

$$p(\mu, \sigma^2) = p(\mu \mid \sigma^2) \cdot p(\sigma^2) \tag{A2.34a}$$

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,

$$\sigma^2 \sim \chi^{-2}_{\nu_0,\sigma_0^2}, \qquad \mu \mid \sigma^2 \sim \mathrm{N}\left(\mu_0, \frac{\sigma^2}{\kappa_O}\right) \tag{A2.34b}$$

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.

The resulting marginal posteriors become

$$\sigma^2 \mid \mathbf{x} \sim \chi^{-2}_{\nu_n,\sigma_n^2}, \quad \text{and} \quad \mu \mid \mathbf{x} \sim \mathrm{t}_{\nu_n}\left(\mu_n, \frac{\sigma_n^2}{\kappa_n}\right) \tag{A2.35}$$

where $\mathrm{t}_n(\mu, \sigma^2)$ denotes a t-distribution with n degrees of freedom, mean μ, and scale parameter σ^2, and

$$\nu_n = \nu_0 + n, \qquad \kappa_n = \kappa_0 + n \tag{A2.36a}$$

$$\mu_n = \mu_0 \frac{\kappa_0}{\kappa_n} + \overline{x}\,\frac{n}{\kappa_n} = \mu_0 \frac{\kappa_0}{\kappa_0 + n} + \overline{x}\,\frac{n}{\kappa_0 + n} \tag{A2.36b}$$

$$\sigma_n^2 = \frac{1}{\nu_n}\left(\nu_0\sigma_0^2 + \sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{\kappa_0 n}{\kappa_n}(\overline{x} - \mu_0)^2\right) \tag{A2.36c}$$

CONJUGATE PRIORS

The use of a prior density that conjugates the likelihood allows for analytic expressions of the posterior density. As we will see in Appendix 3, this is critical in developing a Gibbs sampler for our problem of interest. Table A2.2 summarizes the conjugate priors for several common likelihood functions, with the various families of distributions discussed below.

Table A2.2. Conjugate priors for common likelihood functions. If one uses the distribution family of the conjugate prior with its paired likelihood function, the resulting posterior is in the same distribution family (albeit, of course, with different parameters) as the prior.

Likelihood	Conjugate prior	Equation
Binomial	Beta	A2.38a
Multinomial	Dirichlet	A2.37a
Poisson	Gamma	A2.27a
Normal		
μ unknown, σ^2 known	Normal	A2.18a
μ known, σ^2 unknown	Inverse Chi-Square	A2.30a
Multivariate Normal		
$\boldsymbol{\mu}$ unknown, $\mathbf{V}$ known	Multivariate Normal	LW 8.24
$\boldsymbol{\mu}$ known, $\mathbf{V}$ unknown	Inverse Wishart	A2.41

The Beta and Dirichlet Distributions

With a binomial, each trail (observation) has two possible outcomes and the likelihood is a function of the sample size (number of trails) n and a single success probability p (as the two outcomes on any given trial have probabilities p and $1-p$). The generalization of this is the multinomial, where now each trail has k possible outcomes, which requires $k-1$ success probabilities to describe the likelihood. In particular, for a total of n observations, the probability that n_1 are in catergory 1, n_2 in catergory 2, $\cdots$, n_k in catergory k is

$$p(n_1, \cdots n_k) = \frac{n!}{n_1!\, n_2! \cdots n_k!}\; p_1^{n_1} \cdots p_k^{n_k}$$

In this case the **Dirichlet distribution** is an appropriate prior. Letting $\mathbf{x} = (x_1, x_2, \cdots, x_k)$ denote the k success probabilites, when $\mathbf{x} \sim$ Dirichlet$(\alpha_1, \cdots, \alpha_k)$,

$$p(x_1, \cdots x_k) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)}\; x_1^{\alpha_1 - 1} \cdots x_k^{\alpha_k - 1} \tag{A2.37a}$$

where

$$\alpha_0 = \sum_{i=1}^{k} \alpha_i, \quad 0 \le x_i < 1, \quad \sum_{i=1}^{k} x_i = 1, \quad \alpha_i > 0 \tag{A2.37b}$$

At first blush, this looks like the multinomial density function (with $\alpha_i - 1 = n_i$). The difference is that the multinomial is over a set of discrete random variables, returning the expected probabilities for any vector of discrete number of counts (successes) in each category. Here the discrete number of counts (the data) are random and the continuous success parameters are fixed. Conversely, the Dirichlet treats an equivalent of the vector of outcomes (generalized to non-integers) as fixed and returns the continuous distribution for all possible configurations of the success parameters given this data, so that the data is fixed, and the success parameters are random. A few key moments of this distribution are

$$\mu_{x_i} = \frac{\alpha_i}{\alpha_0}, \quad \sigma^2(x_i) = \frac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}, \quad \sigma^2(x_i, x_j) = -\frac{\alpha_i\, \alpha_j}{\alpha_0^2(\alpha_0 + 1)} \tag{A2.37c}$$

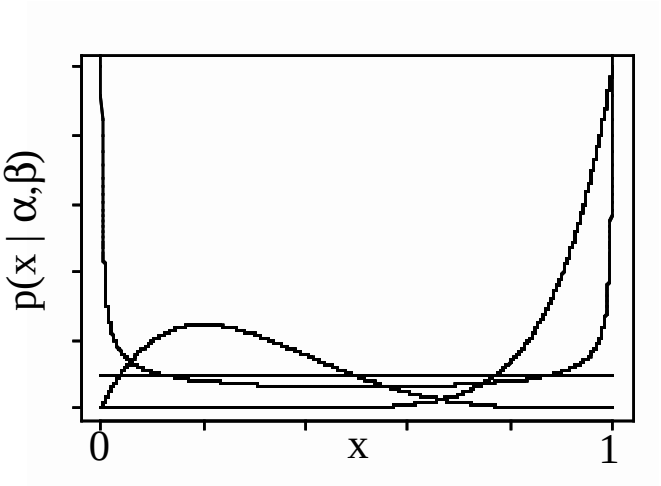

Figure A2.3. For $\alpha = \beta = 1$, the beta distribution is just the uniform distribution over $(0, 1)$. It can also be U-shaped ($\alpha = \beta = 0.5$ in the figure), unimodal ($\alpha = 2, \beta = 5$), or L-shaped ($\alpha = 10, \beta = 1$). Since it symmetric in α and β, switching their parameter values generates a distribution of the same shape translated about 0.5.

An important special case of the Dirichlet (for $k = 2$ classes) is the **Beta distribution**,

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\; x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for} \quad 0 \le x \le 1, \quad \alpha, \beta > 0 \tag{A2.38a}$$

which has mean and variance become

$$\mu = \frac{\alpha}{\alpha+\beta}, \qquad \sigma^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta-1)} \tag{A2.38b}$$

As Figure A2.3 illustrates, the beta distribution is extremely flexible, and can be flat, unimodal, U- or L-shaped, depending on the choice of α and β.

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 vectors 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 resulting random ($k \times k$ symmetric, positive definite) sample covariance matrix

$$\mathbf{W} = \sum_{i=1}^{n} \mathbf{x}_i\, \mathbf{x}_i^T \sim \text{W}_n(\mathbf{V}) \tag{A2.39}$$

Hence, the sum follows a Wishart with n degrees of freedom and parameter $\mathbf{V}$. Recalling that the sum of n squared unit normals follows a χ^2_n distribution, the Wishart is the natural extension to the multivariate normal. Indeed, for $k = 1$ with $\mathbf{V} = (1)$, the Wishart is just a χ^2_n 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, Chapter 11). The probability density function for a Wishart is given by

$$p(\mathbf{W}) = 2^{-nk/2}\pi^{-k(k-1)/k}\, |\, \mathbf{V}\, |^{-n/2}\, |\, \mathbf{W}\, |^{(n+k+1)/2}\, \frac{\exp\left(-\frac{1}{2}\text{tr}\left[\,\mathbf{V}^{-1}\mathbf{W}\,\right]\right)}{\prod_{i=1}^{k}\Gamma\left(\frac{n+1-i}{2}\right)} \tag{A2.40}$$

The **trace** (tr) of a matrix is just the sum of its diagonal elements, e.g., $\text{tr}(\mathbf{A}) = \sum A_{ii}$. 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**. Odell and Feiveson (1966) present an algorithm to obtain generate random draws from the Wishart. The density function for an Inverse-Wishart distributed random matrix $\mathbf{W}$ is

$$p(\mathbf{W}) = 2^{-nk/2}\pi^{-k(k-1)/k} \mid \mathbf{V} \mid^{n/2} \mid \mathbf{W} \mid^{-(n+k+1)/2} \frac{\exp\left(-\frac{1}{2}\text{tr}\left[\mathbf{V}\mathbf{W}^{-1}\right]\right)}{\prod_{i=1}^{k}\Gamma\left(\frac{n+1-i}{2}\right)} \tag{A2.41}$$

which is the distribution of the inverse of the sample covariance matrix.

References

Berger, J. O. 1985. *Statistical decision theory and Bayesian analysis.* Springer-Verlag, New York. [A2]

Carlin, B. P., and T. A. Louis. 2000. *Bayes and empirical Bayes methods for data analysis,* 2nd Ed. Chapman and Hall. [A2]

Efron, B. 1986. Why isn't everyone a Bayesian? *Amer. Stat.* 40: 1-11. [A2]

Gelman, A., J. B. Carlon, H. S. Stern, and D. B. Rubin. 2013. *Bayesian data analysis*, 3rd edition. Chapman and Hall, New York. [A2]

Glymour, C. 1981. Why I am not a Bayesian, *In* C. Glymour D. (ed.), *Theory and evidence*, pp. 63–93. University of Chicago Press, Chicago. [A2]

Jeffreys, H. S. 1961. *Theory of probability*, 3rd ed. Oxford University Press. [A2]

Kendall, M. G., and W. R. Buckland. 1971. *A dictionary of statistical terms.* Hafner, New York. [A2]

Lee, P. M. 2013. *Bayesian statistics: An introduction*, 4th ed. Wiley, New York. [A2]

Lindley, D. V. 1965. *Introduction to probability and statistics from a Bayesian viewpoint* (2 Volumes), University Press, Cambridge. [A2]

Merton, R. K. 1965. *On the shoulders of giants*. Free Press, New York. [A2]

Odell, P. L., and A. H. Feiveson. 1966. A numerical procedure to generate a sample covariance matrix. *Amer. Stat. Assoc. Journ.* 61: 198–203. [A2]

O'Hara, R. B., J. M. Cano, O. Ovaskainen, C. Teplitsky, and J. S. Alho. 2008. Bayesian approaches in evolutionary quantitative genetics. *J. Evol Biol.* 21: 949–957. [20]

Sorensen, D. and D. Gianola. 2002. *Likelihood, Bayesian and MCMC methods in quantitative genetics.* Springer. [A2, A3]

Stigler, S. M. 1980. Stigler's Law of Eponymy. *Trans. NY Acad. Sci., Ser 2* 39: 147–158. [A2]

Stigler, S. M. 1983. Who discovered Bayes's theorem? *Amer. Stat.n* 37: 290–296. [A2]

Van Dongen, S. 2006. Prior specification in Bayesian statistics: three cautionary tales. *J. Theor. Biol.* 242: 90–100. [20]