A Compendium of Conjugate Priors

Daniel Fink
Environmental Statistics Group
Department of Biology
Montana State Univeristy
Bozeman, MT 59717

May 1997

Abstract

This report reviews conjugate priors and priors closed under sampling for a variety of data generating processes where the prior distributions are univariate, bivariate, and multivariate. The effects of transformations on conjugate prior relationships are considered and cases where conjugate prior relationships can be applied under transformations are identified. Univariate and bivariate prior relationships are verified using Monte Carlo methods.

Contents

1 Introduction

Experimenters are often in the position of having had collected some data from which they desire to make inferences about the process that produced that data. Bayes' theorem provides an appealing approach to solving such inference problems. Bayes theorem,

$$g(\theta \mid x_1, \ldots, x_n) = \frac{\pi(\theta)\ L(\theta \mid x_1, \ldots, x_n)}{\int \pi(\theta)\ L(\theta \mid x_1, \ldots, x_n) d\theta} \tag{1}$$

is commonly interpreted in the following way. We want to make some sort of inference on the unknown parameter(s), θ, based on our prior knowledge of θ and the data collected, $x_1, \ldots, x_n$. Our prior knowledge is encapsulated by the probability distribution on θ, $\pi(\theta)$. The data that has been collected is combined with our prior through the likelihood function, $L(\theta \mid x_1, \ldots, x_n)$. The normalized product of these two components yields a probability distribution of θ conditional on the data. This distribution, $g(\theta \mid x_1, \ldots, x_n)$, is known as the posterior distribution of θ. Bayes' theorem is easily extended to cases where is θ multivariate, a vector of parameters. Immediately noticeable in (**??**), unlike most classical methods, is the direct way in which prior information is incorporated into the inference problem. O'Hagan (1994) discusses the merits of using prior information in inference.

Although Bayes' theorem, the cornerstone of Bayesian Statistics, is mathematically simple, its implementation can prove troublesome. The difficulties lie in the normalizing constant, the denominator of (1.1). The product of the prior and likelihood functions must be integrated over the valid domain of the parameter(s) being estimated. This poses the central practical problem: finding analytical and/or numerical solutions for the integral. Since it is not always possible to find an analytically tractable solution to a well defined integration problem, it is not guaranteed that the product of a valid prior distribution and likelihood function are integrable.

Historically, two solutions to these integration problems have been sought. Before the widespread availability of computers, research centered on deriving pairs of likelihood functions and prior distributions with convenient mathematical properties including tractable analytic solutions to the integral. These families of prior distributions are known as conjugate priors or natural conjugate priors. The most recent work has focused on numerical methods that rely on the availability of cheap and powerful computers. These approaches circumvent the analytical integration problem by computing numerical approximations to the integrals needed. The drawback of numerical methods is that they are computationally expensive sometimes requiring large amounts of supercomputer time (and some are simply impractical on any existing computer). This difference in the computational cost of the two approaches increases dramatically with the dimensionality of the prior distribution. The differences in the two approaches suggests two roles for which the conjugate priors are best suited. First, conjugate priors provide a straightforward way to verify any numerical approximation procedure. Second, conjugate priors are the only alternative to numerical methods in settings where the high dimensionality of the prior renders numerical methods computationally infeasible.

1.1 Defining Conjugate Prior Families

With any problem of statistical inference it is the definition of the problem itself that determines the type of data involved. And the analyst's understanding of the process from which the data arise determines the appropriate likelihood function to be used with Bayes' theorem. Thus, our only avenue to produce an analytically tractable solution to the integral is through the choice of the prior given the likelihood function. In the search for a family of conjugate prior distributions we must remember that we are not at liberty to simply choose any prior distribution that works mathematically; we must specify a conjugate prior distribution that adequately describes the experimenter's knowledge of the unknown parameter(s) before the experiment is executed. Consequently, to be of practical use, a conjugate prior family of distributions must produce an analytically tractable solution to the integration problem and it must be flexible enough to model our prior degree of belief in the parameters of interest.

Conjugate prior families were first discussed and formalized by Raiffa and Schlaifer (1961). The definition and construction of conjugate prior distributions depends on the existence and identification of sufficient statistics of fixed dimension for the given likelihood function. If such sufficient statistics exist for a given likelihood function then they permit the dimensionality of the data handled by Bayes' theorem to be reduced. A data set of an arbitrary size can be completely characterized by a fixed number of summaries with respect to the likelihood function. It can be shown that when there exists a set of fixed-dimension sufficient statistics there must exist a conjugate prior family (Raiffa and Schlaifer 1961 and DeGroot 1970).

A conjugate prior is constructed by first factoring the likelihood function into two parts. One factor must be independent of the parameter(s) of interest but may be dependent on the data. The second factor is a function dependent on the parameter(s) of interest and dependent on the data only *through* the sufficient statistics. The conjugate prior family is defined to be proportional to this second factor. Raiffa and Schlaifer (1961) also show that the posterior distribution arising from the conjugate prior is itself a member of the same family as the conjugate prior. When the prior and posterior distributions both belong to the same family, the prior is said to be closed under sampling. Furthermore, because the data is incorporated into the posterior distribution only through the sufficient statistics, there will exist relatively simple formulas for updating the prior into the posterior. These results constitute a constructive definition for conjugate prior families that have attractive mathematical properties that go beyond analytical tractability. The construction of a conjugate prior and the associated properties are demonstrated with a simple example.

Suppose that we are given data that is known to be independent and identically distributed from a normal process with known variance and unknown mean. We wish to infer the mean of this process. The likelihood of a single data point from this process with a given value of the mean, μ, is equal to the probability of drawing the data point from a normal distribution:

$$f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{2}$$

where σ^2 is the known variance of the process. Since the data is known to be independent, the probability of drawing the whole data set is equal to the product of the probabilities of drawing each individual data point,

$$f(x_1, \ldots, x_n \mid \mu) = f(x_1 \mid \mu) \ldots f(x_n \mid \mu). \tag{3}$$

If we substitute equation (1.2) for $f(x \mid \mu)$ and simplify, the likelihood of the data is

$$L(\mu \mid x_1, \ldots, x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right). \tag{4}$$

We can further simplify the argument of the exponent in (1.4) by using the identity

$$\sum_{i=1}^{n}(x_i-\mu)^2 = n\left(\mu-\overline{x}\right) + \sum_{i=1}^{n}(x_i-\overline{x})^2. \tag{5}$$

Substitution of (1.5) into (1.4) yields

$$L(\mu \mid x_1, \ldots, x_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left(-\frac{n\left(\mu-\overline{x}\right)}{2\sigma^2} - \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{2\sigma^2}\right). \tag{6}$$

The factor of the likelihood function that is dependent on the parameter of interest, μ , and dependent on the data only through some sufficient statistic(s) is the single factor of equation (1.6) that is a function of , μ,

$$\exp\left(-\frac{n\left(\mu-\overline{x}\right)}{2\sigma^2}\right). \tag{7}$$

Inspection of Equation 1.7 reveals that the sufficient statistics are $\overline{x}$ and n. Therefore, all of the information that we need to know about the data set is contained in the number of data points in that set and their mean. Furthermore, Equation 1.7 is itself proportional to a normal distribution of μ. Thus, the family of conjugate priors for this data generating process must be the family of normal distributions. Now that we have identified the conjugate prior family we will derive the formulas that update the prior into the posterior distribution which will demonstrate closure under sampling.

Once the family of conjugate priors is known one must specify the unique member of that family that best represents the prior information. This involves finding specific values for the parameter(s) that define the conjugate distribution itself. To avoid confusion with the parameter(s) that we want to make inferences about, the parameter(s) that index the conjugate family are called hyperparameter(s). Generally, each hyperparameter must be specified by some type of prior summary. Subsequently, there is a trade-off between the complexity of the conjugate prior family and the mathematical ease of using the prior distribution. If we let the mean and variance hyperparameters of the prior be m and v respectively, then the prior distribution is

$$\pi\left(\mu \mid m, v\right) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{\left(\mu - m\right)^2}{2v}\right). \tag{8}$$

The factor of the posterior distribution that is proportional to μ must be proportional to the product of the factors of the prior and the likelihood functions that are themselves proportional to μ. Therefore, the posterior distribution is proportional to

$$\exp\left(-\frac{n\left(\mu - \overline{x}\right)}{2\sigma^2}\right) \exp\left(-\frac{\left(\mu - m\right)^2}{2v}\right). \tag{9}$$

If we complete the square on the variable μ for the sum of the arguments of the exponential functions in (1.9) we obtain a single term that is a function of μ. Using this term, the posterior distribution proportional to μ is

$$\pi\left(\mu \mid m', v'\right) \propto \exp\left(-\frac{\left(\mu - \frac{\sigma^2 m + vn\overline{x}}{\sigma^2 + vn}\right)^2}{\frac{2v\sigma^2}{\sigma^2 + vn}}\right). \tag{10}$$

Equation 1.10 is seen to be proportional to a normal distribution with mean and variance

$$m' = \frac{\sigma^2 m + vn\overline{x}}{\sigma^2 + vn} \qquad v' = \frac{v\sigma^2}{\sigma^2 + vn}. \tag{11}$$

The prior is closed under sampling and there are two simple formulas that update the hyperparameters of the prior into those of the posterior.

Raiffa & Schlaifer (1961) and DeGroot (1970) note that if, and only if, the data generating process is a member of the exponential family, and meet certain regularity conditions, must there exist a set of sufficient statistics of fixed dimension. It is only for these processes that there must exist a set of simple operations for updating the prior into the posterior. These results are based on a theorem independently derived by Darmois (1935), Koopman (1936), and Pitman (1936).

1.2 Transformations and the Conjugate Prior

The usefulness of conjugate prior relationships can be greatly extended by noting that these relationships hold under several types of transformations. In the context of the Bayesian inference problem, transformations may be classified according to weather the transformation is a function of only the data, only the unknown parameter(s) of interest, or of both the data and the unknown parameter(s) of interest. We call these three types of transformations process transformations, process reparameterizations, and mixed transformations, respectively. Transformations that introduce additional unknown parameter(s) are not considered in this framework because they essentially define new processes of higher prior dimensionality. For each of the three types of transformations we consider how conjugate prior relationships may be used under the given transformation.

1.2.1 The Effects of Process Transformations

When the transformation is a function of only the data the conjugate prior of the transformed process is the same as the conjugate prior of the untransformed process. The proof of this result arises from the definition of the conjugate prior; the conjugate prior is defined as the family of distributions proportional to the factor of the likelihood function that is itself a function of the parameter(s) of interest and the data as represented *through* the sufficient statistics. We proceed to show that this factor of the transformed likelihood function is proportional to the corresponding factor of the untransformed likelihood function, and, thus, the conjugate prior itself.

Suppose that we have data $x_1, ..., x_n$ that arise independent and identically distributed from a process defined by $f(x \mid \theta)$,where θ may be a vector of unknown parameters that we are interested in making some sort of inference on. For this process we assume that there exists a family of conjugate prior distributions, $g(\theta \mid \phi)$, whose members are indexed by hyperparameter(s), ϕ. The hyperparameter notation will not affect the remainder of this discussion and will be suppressed. Since the conjugate prior exists it is possible to factor the likelihood function, $L(\theta \mid x_1, ..., x_n)$, as stated above:

$$L\left(\theta \mid x_1, ..., x_n\right) = u\left(x_1, ..., x_n\right)\ v\left(T\left(x_1, ..., x_n\right), \theta\right), \tag{12}$$

where $u(\cdot)$ is a function of the data alone, and $v\left(T\left(x_1, ..., x_n\right), \theta\right)$ is a function of the parameter(s) of interest and the data through the sufficient statistics, $T(\cdot)$. This second factor is known as the kernel function. The conjugate prior is defined proportional to the kernel function,

$$g(\theta) \propto v\left(T\left(x_1, ..., x_n\right), \theta\right). \tag{13}$$

Finally, we assume that the transformation that we are interested in is a one-to-one function of the data and any known parameters. That is, the transformation, for the purposes of this discussion can not be a function of any unknown parameters; it can not introduce, for example, an unknown location parameter, and it can not depend on any θ. Let the transformed data $y_1, ..., y_n$ be related to the original data by

$$y_i = h\left(x_i\right). \tag{14}$$

Because the transformation $h\left(x\right)$ is a one-to-one function, its inverse, $h^{-1}\left(x\right)$, exists.

In the case were the process is continuous, for a single transformed data point the transformed likelihood is

$$L(\theta \mid y) = f\left(h^{-1}(y) \mid \theta\right) \left| \frac{d}{dy} h^{-1}(y) \right|, \tag{15}$$

and for the transformed data set the likelihood function is

$$L(\theta \mid y_1, ..., y_n) = f\left(h^{-1}(y_1), ..., h^{-1}(y_n) \mid \theta\right) |J|, \tag{16}$$

where J is the Jacobian. Since the transformed data $y_1, ..., y_n$ is independently distributed the Jacobian simplifies to

$$J = \prod_{i=1}^{n} \frac{d}{dy} h^{-1}(y_i).$$

Recognizing that the right hand side of (1.16) viewed as a function of θ is equal to the untransformed likelihood function of (1.12) times the Jacobian we can rewrite (1.16) as

$$L(\theta \mid y_1, ..., y_n) = u\left(h^{-1}(y_1), ..., h^{-1}(y_n)\right) \; v\left(T\left(h^{-1}(y_1), ..., h^{-1}(y_n)\right), \theta\right) \; |J|. \tag{17}$$

Noting that $x_i = h^{-1}(h(x_i))$, we can rewrite (1.17)

$$L(\theta \mid y_1, ..., y_n) = u(x_1, ..., x_n) \; v(T(x_1, ..., x_n), \theta) \; |J|. \tag{18}$$

Since $|J|$ is also a function of $x_1, ..., x_n$, from our definition of the process transformation, we let

$$\mu(x_1, ..., x_n) = u(x_1, ..., x_n) \; |J|. \tag{19}$$

Substituting (1.19) into (1.18) yields

$$L(\theta \mid y_1, ..., y_n) = \mu(x_1, ..., x_n) \; v(T(x_1, ..., x_n), \theta). \tag{20}$$

This final expression for the transformed likelihood is the product of a function of the data, $\mu(\cdot)$, and the kernel function of the untransformed likelihood. Consequently, $g(\theta)$,the conjugate prior of the untransformed process is also the conjugate prior for the transformed process. This proof is easily modified for discrete likelihood functions by letting all Jacobian terms equal 1. Furthermore, this result makes intuitive sense because it conforms to the likelihood principle. That is, given a particular data set it is plain to see that the information extracted from that data with respect to the parameters of interest, θ, is the same whether or not a process transformation is employed.

If one is considering a many-to-one process transformation, similar results can be obtained. If it is possible to partition the domain of the transformation into k disjoint subsets such that $h_j(x), \quad j = 1, ..., k$, is one-to-one over each subset then the transformed likelihood for a single data point becomes,

$$L(\theta \mid y) = \sum_{j=1}^{k} f\left(h_j^{-1}(y) \mid \theta\right) \left| \frac{d}{dy} h_j^{-1}(y) \right|. \tag{21}$$

There are now k terms in the transformed likelihood function, each of which can be factored into (1.20). Accordingly, the conjugate prior, $g^* (\theta)$, for the transformed likelihood becomes a *mixture* of k conjugate priors:

$$g^* (\theta) = \sum_{j=1}^{k} a_j g_j (\theta), \tag{22}$$

$$where \qquad \sum_{j=1}^{k} a_j = 1. \tag{23}$$

Conjugate prior mixtures are themselves conjugate priors; they produce posterior distributions that are updated mixtures of the updated conjugate prior components. The updating process and properties of conjugate priors mixtures are discussed by O'Hagan(1994), Diaconis & Ylvsaker(1985), and Dalal & Hall(1983). Because many-to-one transformations produce multiple termed likelihood functions, the number of terms in the associated conjugate prior mixture increases with each data point analyzed. Therefore, it is advisable to consider sequential updating for analyzing data sets in this situation.

1.2.2 The Effects of Process Reparameterizations

Process reparameterizations are transformations that act on only the parameters of interest, θ. If one considers a process reparameterization that is one-to-one then the conjugate prior of the transformed parameters of interest is simply the transformation of original conjugate prior. For example, Raiffa and Schlaifer (1961) parameterize the normal process' unknown dispersion in terms of the precision and prove that the resulting conjugate prior is a gamma distribution. La Valle (1970) parameterizes the unknown dispersion in terms of the standard deviation and proves that the resulting conjugate prior is a 'inverted half gamma' distribution. If the precision is distributed as a gamma distribution, then it is straightforward to show that the square root of the reciprocal of the precision, the standard deviation, is itself distributed as an inverted half gamma. Similarly, one finds other parameterizations of the same process parameters in the literature. Box and Tiao (1973) use the inverted χ_2 and inverted χ conjugate prior distributions for the unknown variance and standard deviation parameters, respectively. These two distributions are less flexible special cases of the inverted gamma and inverted half gamma distributions

Similar to the many-to-one process transformations, many-to-one process reparameterizations yield conjugate prior mixtures. Again, if it is possible to partition the domain of the transformation, the parameter space, into k disjoint subsets such that $\psi = w_j (\theta)$, $\quad j = 1, ..., k$, is one-to-one over each subset then the transformed likelihood function becomes

$$L (\psi \mid x_1, ..., x_n) = \sum_{j=1}^{k} L \left(w_j^{-1} (\psi) \mid x_1, ..., x_n \right) |J_j| . \tag{24}$$

where the general form of the Jacobian, assuming l elements in θ and ψ, is

$$J_j = \begin{vmatrix} \frac{\delta\theta_1}{\delta\psi_1}\left(w_j^{-1}(\psi)\right) & \cdots & \frac{\delta\theta_1}{\delta\psi_l}\left(w_j^{-1}(\psi)\right) \\ \vdots & \ddots & \vdots \\ \frac{\delta\theta_l}{\delta\psi_1}\left(w_j^{-1}(\psi)\right) & \cdots & \frac{\delta\theta_l}{\delta\psi_l}\left(w_j^{-1}(\psi)\right) \end{vmatrix}. \tag{25}$$

We recognize that the each of the k terms of the transformed likelihood is itself proportional to the conjugate prior transformed by the associated transformation, that is,

$$L\left(w_j^{-1}(\psi) \mid x_1, ..., x_n\right)|J_j| \propto g(w_j^{-1}(\psi))\,|J_j|\,. \tag{26}$$

Therefore, the conjugate prior of a many-to-one process reparameterization is a mixture of the transformed conjugate priors (1.26).

It is interesting to note that many-to-one process transformations and reparameterizations result in conjugate prior mixtures made up of the conjugate priors of the associated untransformed processes. The many-to-one transformations in both cases introduce another source of uncertainty into these processes. This uncertainty takes form as a *multi-termed* likelihood function and it is carried through to the conjugate prior mixture, where the mixture weights constitute k new unknown parameters that are updated by the Bayes' Theorem.

1.2.3 Mixed Transformations

Mixed transformations are functions of both the data and the unknown parameter(s) of interest. The Jacobian that results from such a transformation will itself be a function of the data and the parameters of interest. This means that the transformed likelihood (1.17) can not generally be factored into the product of a function of the data, and the untransformed likelihood. Therefore, it will only be special cases where the conjugate prior under the mixed transformation belongs to the same family as the conjugate prior associated with the untransformed process.

1.3 Scope, Organization, and Verification of the Compendium

There has been little recent work to expand or update the collection of known families of conjugate prior distributions. Even the most comprehensive lists of conjugate prior relationships (DeGroot(1970) and LaValle(1970)) offer only limited amendments to Raiffa & Schlaifer's (1961) work. This report augments the list of conjugate priors for univariate and multivariate data generating processes. Additionally, this report presents non-conjugate priors that exhibit closure under sampling for some common data generating processes. A short list of Jeffreys's priors that lead to analytically tractable posterior distributions is also presented. Attempts have been made to standardize some of the notational differences and correct some of the mistakes currently found in the literature.

It should be pointed out that much of the recent literature on conjugate relationships has been based on an alternative characterization of the conjugate prior. Diaconis and Ylvisaker(1979) show that for a natural exponential family, the conjugate prior family of a canonical parameter is characterized by the property that the posterior expected value of the mean parameter is linear in the canonical sufficient statistic. The purpose of this report is to compile those prior/likelihood pairs that yield analytically tractable and 'convenient' posteriors. We have decided to focus our attention on conjugate prior relationships, and when the conjugate prior is not obtainable, priors closed under sampling were sought. Raiffa & Schlaiffer's (1961) conjugate prior definition is the best tool suited to this task:

1. it is constructive in nature,
2. it is not limited to processes that are members of the exponential family,
3. it allows one to work with the 'common' parameterization of the data generating processes, and
4. it is easy to adapt its use to find priors that are at least closed under sampling.

This list is organized according to the dimensionality of the parameters of interest and the data type. The conjugate priors of processes that are univariate in the parameter space are presented, bivariate in the parameter space, multivariate data generating processes, and then Jeffreys's priors. Within each of these categories, discrete data generating processes are discussed first. For each conjugate prior relationship the explicit form of the conjugate prior, the likelihood function, and the relevant sufficient statistics are presented. Formulas for updating the hyperparameters are presented and for univariate processes they are verified.

This verification consists of comparing the analytic conjugate posterior distributions to numerically approximated posterior distributions. The numerical procedure used to approximate the posterior distributions is based on the sampling scheme of Smith and Gelfand (1992). This procedure works by sampling parameter values from the prior distribution and weighing each sampled prior value by its likelihood. These likelihood weights are then accumulated in bins and plotted as a histogram to approximate the posterior distribution.

Statistical "Goodness of Fit" tests can not be used to compare the approximate and analytical posterior distributions. The numerical posterior is computed from a monte carlo sampling procedure and its sample size is, by definition, very large, and, the proportion of samples contributing to the tails of any of these posterior distributions is, by definition, relatively small. Therefore, the numerical posterior will very closely approximate the true posterior where the true posterior has high density, and it will produce a relatively poor approximation of the true posterior tails. If one conducted a "Goodness of Fit" test to determine if the two posteriors were the same they would find that because of the large sample size even slight deviations between the tails of the two distributions will lead to the conclusion that the distributions are not the same. Consequently, a deliberate decision has been made not to perform any statistical "Goodness of Fit" tests comparing the two posterior distributions.

The comparison that is done is to determine if the approximate posterior distributions are converging to the analytical conjugate posterior distributions. To make this comparison of the posterior distributions easier, we only compute and plot the marginals of the posterior distributions. We begin by computing each approximate marginal posterior twice, each time with a different random number seed. The composite of the two resulting histograms is plotted with the difference between each corresponding bar filled. Over this composite histogram we plot the analytical conjugate marginal posterior. Inspection reveals any substantial systematic divergence or deviation between the approximations and the analytical conjugate distributions.

The number of samples from the prior used to generate the histogram approximation of the posterior is determined by finding the smallest number of samples such that the histogram representations of the posterior are stable for a given number of bins. We choose to use 100 bins for the plotted histograms so that the approximations are fairly detailed. The numerical procedure is judged to be stable for a given number of bins and samples when successively produced histograms are similar.

2 Univariate Data, Univariate Priors

2.1 Discrete Data: Bernoulli Process

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Bernoulli process where p, the success of a hit, is unknown. The value of each x_i is either 0 or 1, a 'miss' or a 'hit' respectively. The likelihood function, $L(p \mid x_1, ...x_n)$, computes a measure of how likely a value of p is for a given data vector. For data from a Bernoulli process the likelihood function, proportional to p, is

$$L(p \mid x_1, ...x_n) \propto p^{{}_{i=1}^{n} x}(1-p)^{n-\sum_{i=1}^{n} x}. \tag{27}$$

The sufficient statistics are n, the number of data points, and ${}_{i=1}^{n}x$, the sum of the data.

In view of the fact that the likelihood function determines the influence that the data have on the posterior distribution, it is prudent to consider if the way in which the data are collected can affect the likelihood function. Data from this process could have been collected in one of three ways:

1. n, the total number of data observed could have been predetermined, or
2. ${}_{i=1}^{n}x$, the sum of hits could have been predetermined, or
3. the experiment was unexpectedly interrupted and neither n or ${}_{i=1}^{n}x$ was predetermined.

These three scenarios differ in how it is decided to stop collecting data. Although the likelihood functions for these processes are not the same, the factors of the likelihood functions proportional to p, the parameter of interest, are all equal to Equation 2.1, leading to the same posterior distribution. This makes intuitive sense since the data really provide the same amount of information about the process regardless of the manner of collection. In general, the likelihood function is invariant to the stopping rule used, regardless of the data generating process, so as long as the stopping rule does not depend on the parameters of interest.

If one views (2.1) as a function of the unknown parameter p, it is proportional to the beta distribution. Thus, the conjugate prior, $\pi(\cdot)$, is a beta distribution with hyperparameters $\alpha > 0$ and $\beta > 0$

$$\pi(p \mid \alpha, \beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} & where \quad 0 < p < 1 \\ 0 & otherwise \end{cases}. \tag{28}$$

The closure property is verified by proving that the posterior distribution of the unknown parameter p, the product of the prior and likelihood functions, is also a member of the beta family. The posterior distribution of p, is the beta distribution, $\pi(p \mid \alpha', \beta')$ where the updated hyperparameters are

$$\alpha' = \alpha +_{i=1}^{n} x \qquad \beta' = \beta + n -_{i=1}^{n} x. \tag{29}$$

The example used to verify this conjugate relationship was constructed with a beta prior specified by hyperparameters $\alpha = 3.0$ and $\beta = 3.0$. The sufficient statistics describing the data are $n = 5$ and $_{i=1}^{n} x = 4$. The conjugate posterior distribution is plotted with two 100 bin histograms produced from the numerical estimate of the posterior distribution in Figure 1. The numerical posteriors were generated with 100,000 samples from the prior distribution.

2.2 Discrete Data: Poisson Process

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Poisson process where μ, the mean, is unknown. The value of each x_i is an integer greater than or equal to zero. For data from a Poisson process the likelihood function, $L(\mu \mid x_1, ...x_n)$, proportional to μ, is

$$L(\mu \mid x_1, ...x_n) \propto \begin{cases} \mu^x \exp(-n\mu) & where \quad x_i > 0 \\ 0 & otherwise \end{cases}. \tag{30}$$

The sufficient statistics are n, the number of data points, and $_{i=1}^{n} x_i$, the sum of the data. This likelihood factor is proportional to a gamma distribution of μ. The conjugate prior, $\pi(\cdot)$, is a gamma distribution with hyperparameters $\alpha > 0$ and $\beta > 0$,

$$\pi(\mu \mid \alpha, \beta) = \begin{cases} \frac{\mu^{\alpha-1} \exp(\frac{-\mu}{\beta})}{\Gamma(\alpha)\beta^{\alpha}} & where \quad \mu > 0 \\ 0 & otherwise \end{cases}. \tag{31}$$

The closure property is verified by showing that the posterior distribution of the unknown parameter is also a member of the gamma family. The posterior distribution of μ, is the gamma distribution, $\pi(\mu \mid \alpha', \beta')$ where the updated hyperparameters are

$$\alpha' = \alpha +_{i=1}^{n} x_i \qquad \beta' = \frac{\beta}{1+n}. \tag{32}$$

The example used to verify this conjugate relationship was constructed with a gamma prior specified by hyperparameters $\alpha = 3.0$ and $\beta = 2.0$. The sufficient statistics describing the data are $n = 10$ and $\sum x = 4.0$. Figure 2 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.3 Discrete Data: Negative Binomial Process

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a negative binomial process with process parameters r and p where r is known ($r > 0$) and p is unknown. For data from a negative binomial process the likelihood function, $L(r, p \mid x_1, ..., x_n)$ proportional to p , is

$$L(r, p \mid x_1, ..., x_n) \propto \begin{cases} p^{rx}(1-p)^{\sum x - rn} & where \quad x_i \geq r \\ 0 & otherwise \end{cases}. \tag{33}$$

The sufficient statistics are n, the number of data points, and $\sum_{i=1}^{n} x_i$, the sum of the data. Equation 2.7 is proportional to a beta distribution of p. Consequently, the conjugate prior, $\pi(\cdot)$, is a beta distribution with hyperparameters $\alpha > 0$ and $\beta > 0$

$$\pi(p \mid \alpha, \beta) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} & where \quad 0 < p < 1 \\ 0 & otherwise \end{cases}. \tag{34}$$

The closure property is verified by showing that the posterior distribution of the unknown parameter p is also a member of the beta family. The posterior distribution of p is the beta distribution, $\pi(p \mid \alpha', \beta')$ with the updated hyperparameters

$$\alpha' = \alpha + rn \qquad \beta' = \beta + \textstyle\sum_{i=1}^{n} x_i - rn. \tag{35}$$

The example used to verify this conjugate relationship was constructed with a beta prior specified by hyperparameters $\alpha = 8.0$ and $\beta = 3.0$ and process parameter r known to be equal to 4. The sufficient statistics describing the data are $n = 10$ and $\sum x = 52$. Figure 3 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.4 Discrete Data: Hypergeometric Process

The conjugate prior for a hypergeometric process is discussed by Dyer and Pierce (1993). Suppose that data x arise from a hypergeometric process where process parameter N, the total population size, is known ($N > 0$) and M, the number of target members of the population, is unknown ($0 \leq M \leq N$). If n objects are sampled without replacement, x is the number of target members in that sample. For data from a hypergeometric process the likelihood function, $L(n, M, N \mid x)$, is

$$L(n, M, N \mid x) = \begin{cases} \frac{MxN-Mn-x}{Nn} & where \quad 0 \leq x_i \leq M \\ 0 & otherwise \end{cases}. \tag{36}$$

Equation 2.10 is proportional to a beta-binomial distribution of M. Consequently, the conjugate prior, $\pi(\cdot)$, is a beta-binomial distribution with hyperparameters $\alpha > 0$ and $\beta > 0$. Using the following notion for a beta function with parameters a and b,

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}, \tag{37}$$

the beta-binomial prior is

$$\pi(M \mid N, \alpha, \beta) = \begin{cases} \frac{NMB(\alpha+M, \beta+N-M)}{B(\alpha,\beta)} & where \quad 0 \leq M \leq N \\ 0 & otherwise \end{cases}. \tag{38}$$

The closure property is verified by showing that the posterior distribution of the unknown parameter M is also a member of the beta-binomial family. The posterior distribution is

$$\pi\left(M \mid N, \alpha, \beta\right)=\begin{cases} \frac{N-nM-xB(\alpha+M,\beta+N-M)}{B(\alpha+x,\beta+n-x)} & where \quad 0 \leq M \leq N \\ 0 & otherwise \end{cases}. \tag{39}$$

which is also a beta-binomial distribution.

The example used to verify this conjugate relationship was constructed with a beta-binomial prior specified by hyperparameters $\alpha = 3.0$, $\beta = 3.0$, and process parameter N known to be equal to 15. The sample size, n, is 7 and $x = 5$. Figure 4 shows the conjugate posterior plotted with the histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.5 Continuous Data: Uniform Process

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a uniform process where w, the upper boundary, is unknown and the lower boundary is known to be 0. The likelihood function, $L(w \mid x_1, ...x_n)$, proportional to w, is

$$L(w \mid x_1, ...x_n)=\begin{cases} \frac{1}{w^n} & where \quad w > \max(x_i) \\ 0 & otherwise \end{cases}. \tag{40}$$

The sufficient statistics are n, the number of data points, and $\max(x_i)$ the value of the maximum data point. Viewed as a function of w, Equation 2.14 is proportional to the Pareto distribution of w. The conjugate prior, $\pi(\cdot)$, is a Pareto distribution with hyperparameters $\alpha > 0$ and $\beta > 0$,

$$\pi\left(w \mid \alpha, \beta\right)=\begin{cases} \frac{\alpha\beta^{\alpha}}{w^{\alpha+1}} & where \quad w > \beta \\ 0 & otherwise \end{cases}. \tag{41}$$

The restrictions necessary for the prior and the likelihood functions to be nonzero allow the posterior distribution to be nonzero only where $w > \max(x_i, \beta)$. When this condition is met, the posterior distribution proportional is a Pareto distribution, $\pi\left(w \mid \alpha', \beta'\right)$,with hyperparameters

$$\alpha' = \alpha + n \qquad \beta' = \max(x_i, \beta). \tag{42}$$

These results are easily applied, with slight modification, to a uniform process where the lower boundary is unknown and the upper boundary is known or when the known endpoint is non-zero.

The example used to verify this conjugate relationship was constructed with a Pareto prior specified by hyperparameters $\alpha = 0.50$ and $\beta = 1.0$. The sufficient statistics describing the data are $n = 10$ and $\max(x_i) = 2.23$. Figure 5 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.6 Continuous Data: Pareto Process

Arnold and Press (1989) study conjugate prior relationships for the Pareto bivariate data generating process. The conjugate priors suggested for the two corresponding univariate priors are presented here.

2.6.1 Unknown Precision Parameter

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Pareto process where α, the process shape parameter, is known and β, the process precision parameter is unknown. The likelihood function, $L(\beta \mid x_1, ..., x_n)$, proportional to β, is

$$L(\beta \mid x_1, ..., x_n) \propto \begin{cases} \beta^{n\alpha} & where \quad \min(x_i) > \beta \\ 0 & otherwise \end{cases} . \tag{43}$$

The sufficient statistic is n, the number of data points. Viewed as a function of β, Equation 2.17 is proportional to a Pareto distribution of β. The conjugate prior, $\pi(\cdot)$, is a Pareto distribution with hyperparameters $a > 0$ and $b > 0$

$$\pi(\beta \mid a, b) = \begin{cases} \frac{ab^a}{\beta^{a+1}} & where \quad \beta > b \\ 0 & otherwise \end{cases} . \tag{44}$$

The posterior distribution proportional to is Equation 2.18 is a Pareto distribution, $\pi(\beta \mid a', b')$, with hyperparameters

$$a' = a - \alpha n \qquad b' = b. \tag{45}$$

It is worth noting again that a' , the posterior shape hyperparameter, must be greater than zero. This imposes the following additional constraint on the prior shape hyperparameter: $a > \alpha n$.

The example used to verify this conjugate relationship was constructed with a Pareto prior specified by hyperparameters $a = 5.0$, $b = 1.5$, and the known process shape parameter, $\alpha = 0.30$. The sufficient statistic describing the data is n =4. Figure 6 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.6.2 Unknown Shape Parameter

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Pareto process where β, the process precision parameter is known and α, the process shape parameter, is unknown. The likelihood function, $L(\alpha \mid x_1, ..., x_n)$, proportional to α, is

$$L(\alpha \mid x_1, ..., x_n) \propto \begin{cases} \frac{\alpha^n \beta^{n\alpha}}{m^{\alpha+1}} & where \quad \min(x_i) > \beta \\ 0 & otherwise \end{cases} \tag{46}$$

$$where \quad m = \prod_{i=1}^{n} x_i.$$

The sufficient statistics are n, the number of data points, and m, the product of the data. Viewed as a function of α, Equation 2.20 is proportional to a gamma distribution. The conjugate prior, $\pi(\cdot)$, is a gamma distribution with hyperparameters $\alpha > 0$ and $\beta > 0$,

$$\pi(\alpha \mid a, b) = \begin{cases} \frac{\alpha^{a-1}\exp(\frac{-\alpha}{b})}{\Gamma(a)b^{a}} & where \quad \alpha > 0 \\ 0 & otherwise \end{cases} . \tag{47}$$

The posterior distribution, $\pi(\alpha \mid a', b')$, is a gamma distribution with hyperparameters

$$a' = a + n \qquad b' = \frac{1}{\frac{1}{b} + \ln(m) - n\ln(\beta)}, \tag{48}$$

proving that there is closure under sampling. However, since the posterior gamma scale parameter, b', must be greater than zero there is one more condition placed on the conjugate relationship:

$$b < (n\ln(\beta) - \ln(m))^{-1}.$$

The example used to verify this conjugate relationship was constructed with a gamma prior specified by hyperparameters $a = 2.0$, $b = 1.5$, and the known process precision parameter $\beta = 0.50$. The sufficient statistics describing the data are $n = 4$ and the product of the data, $m = 10.23$. Figure 7 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.7 Continuous Data: Exponential Process

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a exponential process where θ, the mean, is unknown. The value of each x_i is a real number greater than zero. The likelihood function, $L(\theta \mid x_1, ..., x_n)$, proportional to θ is

$$L(\theta \mid x_1, ..., x_n) \propto \begin{cases} \theta^{n}\exp\left(-\theta \sum_{i=1}^{n} x_i\right) & where \quad x_i > 0 \\ 0 & otherwise \end{cases} . \tag{49}$$

The sufficient statistics are n, the number of data points, and $\sum_{i=1}^{n} x$, the sum of the data. Equation 2.23 is proportional to a gamma distribution of θ. The conjugate prior, $\pi(\cdot)$, is a gamma distribution with hyperparameters $\alpha > 0$ and $\beta > 0$

$$\pi(\theta \mid \alpha, \beta) = \begin{cases} \frac{\theta^{\alpha-1}\exp(\frac{-\theta}{\beta})}{\Gamma(\alpha)\beta^{\alpha}} & where \quad \theta > 0 \\ 0 & otherwise \end{cases} . \tag{50}$$

The posterior distribution, $\pi(\theta \mid \alpha', \beta')$, is a gamma distribution with hyperparameters

$$\alpha' = \alpha + n \qquad \beta' = \frac{\beta}{1 + \beta\sum_{i=1}^{n} x_i}. \tag{51}$$

The example used to verify this conjugate relationship was constructed with a gamma prior specified by hyperparameters $\alpha = 2.0$ and $\beta = 1.0$. The sufficient statistics describing the data are $n = 10$ and $\sum x = 2.0$. Figure 8 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.8 Continuous Data: Gamma Process

The gamma process is a bivariate process whose two parameters are found either as shape and scale parameters or their reciprocals. Each case considered here is made univariate in the unknown parameter by assuming one or another of the parameters is known.

2.8.1 Unknown Rate Parameter

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a gamma process where α, the process shape parameter, is known and $\frac{1}{\beta}$, the process rate parameter (or the reciprocal of the process scale parameter) is unknown. The likelihood function, $L(\beta \mid x_1, ..., x_n)$, proportional to β is

$$L(\beta \mid x_1, ..., x_n) \propto \begin{cases} \frac{\exp\left(\frac{-\sum_{i=1}^{n} x_i}{\beta}\right)}{\beta^{\alpha n}} & where \quad x_i > 0 \\ 0 & otherwise \end{cases} . \tag{52}$$

The sufficient statistics are n, the number of data points, and $\sum_{i=1}^{n} x$, the sum of the data. Equation 2.26 is proportional to a gamma distribution of $1/\beta$. The conjugate prior, $\pi(\cdot)$, is a gamma distribution of $1/\beta$ with hyperparameters $a > 0$ and $b > 0$

$$\pi\left(\frac{1}{\beta} \mid a, b\right) = \begin{cases} \frac{\beta^{1-a} \exp\left(\frac{-1}{\beta b}\right)}{\Gamma(a) b^a} & where \quad \frac{1}{\beta} > 0 \\ 0 & otherwise \end{cases} . \tag{53}$$

The posterior distribution, $\pi\left(\frac{1}{\beta} \mid a', b'\right)$, is a gamma distribution with hyperparameters

$$a' = \alpha n + a \qquad b' = \frac{b}{1 + b\sum_{i=1}^{n} x_i}. \tag{54}$$

The example used to verify this conjugate relationship was constructed with a gamma prior specified by hyperparameters $a = 2.0$, $b = 2.0$, and the known process scale parameter $\alpha = 1.54$. The sufficient statistics describing the data are $n = 5$ and $\sum x = 11.34$. Figure 9 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.8.2 Unknown Shape Parameter

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a gamma process where β, the process scale parameter, is known and α, the process shape parameter is unknown. The likelihood function, $L(\alpha \mid x_1, ..., x_n)$, proportional to α is

$$L(\alpha \mid x_1, ..., x_n) \propto \begin{cases} \frac{p^{\alpha-1}}{\beta^{-\alpha n}\Gamma(\alpha)^n} & where \quad x_i > 0 \\ 0 & otherwise \end{cases} , \tag{55}$$

$$where \qquad p = \prod_{i=1}^{n} x_i.$$

The sufficient statistics are n, the number of data points, and p, the product of the data. Equation 2.29 can itself be viewed as proportional to a distribution of α. The conjugate prior, $\pi(\cdot)$, is

$$\pi(\alpha \mid a, b, c) = \begin{cases} \frac{1}{K} \frac{a^{\alpha-1}\beta^{\alpha c}}{\Gamma(\alpha)^b} & where \quad \alpha > 0 \\ 0 & otherwise \end{cases} \tag{56}$$

$$where \qquad K = \int_0^\infty \frac{a^{\alpha-1}\beta^{\alpha c}}{\Gamma(\alpha)^b} d\alpha,$$

and the hyperparameters $a, b, c > 0$. The posterior distribution is specified by (2.29) with the updated hyperparameters

$$a' = ap \qquad b' = b + n \qquad c' = c + n. \tag{57}$$

The example used to verify this conjugate relationship was constructed with a prior specified by hyperparameters $a = 1.5$, $b = 1.0$, and $c = 1.25$ and the known process scale parameter $\beta = 2.0$. The sufficient statistics describing the data are $n = 3$ and $p = 2.213$. Figure 10 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.9 Continuous Data: Normal Process

The normal process is a bivariate process that is most commonly specified in terms of its mean and variance. The two corresponding univariate processes are presented here in terms of their means and precisions, the reciprocal of the variance. This alternative parameterization admits the most elegant conjugate prior distributions and formulas.

2.9.1 Unknown Mean Parameter

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a normal process where the mean, μ, is unknown and the precision, ρ, is known. The likelihood function, $L(\mu \mid x_1, ..., x_n)$, proportional to μ is

$$L(\mu \mid x_1, ..., x_n) \propto \exp\left(-\frac{n\rho}{2}(\mu - \overline{x})^2\right). \tag{58}$$

The sufficient statistics are n, the number of data points, and $\overline{x}$,the mean of the data. As a function of μ, Equation 2.32 is proportional to a normal distribution of μ. The conjugate prior, $\pi(\cdot)$, is a normal distribution with hyperparameters m and $p > 0$

$$\pi(\mu \mid m, p) = \sqrt{\frac{p}{2\pi}} \exp\left(-\frac{p}{2}(\mu - m)^2\right). \tag{59}$$

The posterior distribution, $\pi\left(\mu \mid m', p'\right)$, is a normal distribution with hyperparameters

$$m' = \frac{mp + n\rho\overline{x}}{p + n\rho} \qquad p' = p + n\rho. \tag{60}$$

The example used to verify this conjugate relationship was constructed with a normal prior specified by hyperparameters $m = 2.0$ and $p = 1.0$ with the known precision $\rho = 2.0$. The sufficient statistics describing the data are $n = 10$ and $\overline{x} = 4.0$. Figure 11 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.9.2 Unknown Precision Parameter

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a normal process where the mean, μ, is known and the precision, ρ, is unknown. The likelihood function, $L(\rho \mid x_1, ..., x_n)$, proportional to ρ is

$$L(\rho \mid x_1, ..., x_n) \propto \rho^{\frac{n}{2}} \exp\left(-\frac{\rho SS}{2}\right), \tag{61}$$

where SS is the sum of the squares of the deviations from the mean of the data. The sufficient statistics are n, the number of data points, and SS the sum of the squares. Viewed as a function ρ, Equation 2.35 is proportional to a gamma distribution. The conjugate prior distribution of the precision, $\pi\left(\cdot\right)$, is a gamma distribution with hyperparameters $\alpha > 0$ and $\beta > 0$

$$\pi\left(\rho \mid \alpha, \beta\right) = \begin{cases} \frac{\rho^{\alpha-1}\exp(\frac{-\rho}{\beta})}{\Gamma(\alpha)\beta^{\alpha}} & where \quad \rho > 0 \\ 0 & otherwise \end{cases}. \tag{62}$$

The posterior distribution, $\pi\left(\rho \mid \alpha', \beta'\right)$, is a gamma distribution with hyperparameters

$$\alpha' = \alpha + \tfrac{n}{2} \qquad \beta' = \beta + \tfrac{SS}{2}. \tag{63}$$

The example used to verify this conjugate relationship was constructed with a gamma prior specified by hyperparameters $\alpha = 3.0$, $\beta = 1.5$, and the known process mean is 2.0. The sufficient statistics describing the data are $n = 20$ and $SS = 38.0$. Figure 12 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.10 Continuous Data: Inverse Gaussian Process

Banerjee and Bhattacharyya(1979) study bivariate conjugate prior relationships for the Inverse Gaussian process. The two corresponding univariate processes presented here are specified in terms of a parameter m, the reciprocal mean, and λ, the coefficient of variation multiplied by the mean. The probability density function for the Inverse Gaussian process is

$$f(x \mid m, \lambda) = \begin{cases} \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda m^2}{2x}\left(x - \frac{1}{m}\right)^2\right) & where \quad x > 0 \\ 0 & otherwise. \end{cases} \tag{64}$$

This parameterization admits the most elegant conjugate prior distributions and formulas.

2.10.1 Unknown Reciprocal Mean Parameter

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Inverse Gaussian process where parameter m is unknown and the parameter λ is known. The likelihood function, $L(m \mid x_1, ..., x_n)$, proportional to m is

$$L(m \mid x_1, ..., x_n) \propto \exp\left(-\frac{\lambda}{2}(m^2\Sigma x - 2nm)\right). \tag{65}$$

The sufficient statistics are n, the number of data points, and Σx,the sum of the data. As a function of m, Equation 2.39 is proportional to a normal distribution of m. The conjugate prior, $\pi(\cdot)$, is a normal distribution with hyperparameters μ and $p > 0$

$$\pi(m \mid \mu, p) = \sqrt{\frac{p}{2\pi}} \exp\left(-\frac{p}{2}(m-\mu)^2\right). \tag{66}$$

The posterior distribution, $\pi(m \mid \mu', p')$, is a normal distribution with hyperparameters

$$\mu' = \frac{\lambda n + \rho\mu}{\lambda\Sigma x + \rho} \qquad \lambda' = \lambda\Sigma x + \rho. \tag{67}$$

The example used to verify this conjugate relationship was constructed with a normal prior specified by hyperparameters $\mu = 4.0$ and $p = 0.50$ with the known $\lambda = 1.16$. The sufficient statistics describing the data are $n = 5$ and $\Sigma x = 2.695$. Figure 13 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.10.2 Unknown Parameter λ

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Inverse Gaussian process where the parameter m is known and the parameter λ is unknown. The likelihood function, $L(\lambda \mid x_1, ..., x_n)$, proportional to λ is

$$L(\lambda \mid x_1, ..., x_n) \propto \lambda^{\frac{n}{2}} \exp\left(-\frac{\lambda}{2}\left(m^2\Sigma x - 2nm + \Sigma\frac{1}{x}\right)\right). \tag{68}$$

The sufficient statistics are n, the number of data points, $\Sigma\frac{1}{x}$,the sum of the reciprocal data, and Σx the sum of the data. Viewed as a function λ, Equation 2.42 is proportional to a gamma distribution. The conjugate prior distribution of the precision, $\pi(\cdot)$, is a gamma distribution with hyperparameters $\alpha > 0$ and $\beta > 0$

$$\pi(\lambda \mid \alpha, \beta) = \begin{cases} \frac{\lambda^{\alpha-1}\exp\left(\frac{-\lambda}{\beta}\right)}{\Gamma(\alpha)\beta^\alpha} & where \quad \lambda > 0 \\ 0 & otherwise \end{cases}. \tag{69}$$

The posterior distribution, $\pi(\lambda \mid \alpha', \beta')$, is a gamma distribution with hyperparameters

$$\alpha' = \alpha + \frac{n}{2} \qquad \beta' = \left(\frac{1}{\beta} + \frac{m^2\Sigma x - 2nm + \Sigma\frac{1}{x}}{2}\right)^{-1}. \tag{70}$$

The example used to verify this conjugate relationship was constructed with a gamma prior specified by hyperparameters $\alpha = 2.0$, $\beta = 3.0$, and the known parameter $m = 4.0$. The sufficient statistics describing the data are $n = 5$, $\Sigma x = 2.695$, and $\Sigma \frac{1}{x} = 9.298$. Figure 14 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.11 Continuous Data: Weibull Process

The Weibull process is specified by two parameters, shape, β, and scale, θ. The two corresponding univariate processes presented here are based on the Weibull probability density function

$$f(x \mid \beta, \theta) = \begin{cases} \frac{\beta}{\theta} x^{\beta-1} \exp\left(-\frac{x^\beta}{\theta}\right) & where \quad x > 0 \\ 0 & otherwise. \end{cases} \tag{71}$$

It should be noted that the prior developed here for the shape parameter is not a true conjugate prior. In a loose sense, it is closed under sampling. The lack of a true conjugate prior is due to the fact that there does not exist a sufficient statistic of fixed dimension with respect to the unknown shape parameter.

2.11.1 Unknown Scale Parameter

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Weibull process where the scale parameter, θ, is unknown and the shape parameter, β, is known. The likelihood function, $L(\theta \mid x_1, ..., x_n)$, proportional to θ is

$$L(\theta \mid x_1, ..., x_n) \propto \theta^{-n} \exp\left(-\frac{\sum x^\beta}{\theta}\right). \tag{72}$$

The sufficient statistics are n, the number of data points, and Σx^β,the sum of the data raised to the βth power. As a function of θ, Equation 2.46 is proportional to an inverted gamma distribution of θ. The conjugate prior, $\pi(\cdot)$, is an inverted gamma distribution with hyperparameters a, $b > 0$

$$\pi(\theta \mid a, b) = \begin{cases} \frac{b^{a-1} \exp(\frac{-b}{\theta})}{\Gamma(a-1)\theta^a} & where \quad \theta > 0 \\ 0 & otherwise. \end{cases} \tag{73}$$

The posterior distribution, $\pi(\theta \mid a', b')$, is a normal distribution with hyperparameters

$$a' = a + n \qquad b' = b + \sum_{i=1}^{n} x_i^\beta. \tag{74}$$

The example used to verify this conjugate relationship was constructed with a inverted gamma prior specified by hyperparameters $a = 2.0$ and $b = 1.0$ with the known shape parameter $\beta = 2.0$. The sufficient statistics describing the data are $n = 5$ and $\Sigma x^\beta = 11.846$. Figure 15 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.11.2 Unknown Shape Parameter

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Weibull process where the scale parameter, θ, is known and the shape parameter, β, is unknown. The likelihood function, $L(\beta \mid x_1, ..., x_n)$, proportional to β is

$$L(\beta \mid x_1, ..., x_n) \propto \left(\frac{\beta}{\theta}\right)^n \exp\left(-\beta \ln(P) - \frac{\sum x^\beta}{\theta}\right). \tag{75}$$

The statistics of interest are n, the number of data points, Σx^β ,the sum of the data raised to the βth power, and $P = \prod x$, the product of the data. Viewed as a function β, Equation 2.49 is conformable with the following prior distribution of the shape parameter, $\pi(\beta \mid a, b, d)$, with hyperparameters $a > 0$, $b > \ln(P)$,and $d > 0$,

$$\pi(\beta \mid a, b, d) = \begin{cases} \frac{1}{K}\beta^a \exp\left(-b\beta - \frac{d^\beta}{\theta}\right) & where \quad \beta > 0 \\ 0 & otherwise \end{cases}, \tag{76}$$

where K is the normalizing constant defined as

$$K = \int_0^\infty \beta^a \exp\left(-b\beta - \frac{d^\beta}{\theta}\right) d\beta.$$

The posterior distribution, $\pi(\beta \mid a', b', d')$, has hyperparameters

$$a' = a + n \qquad b' = b - \ln(P) \qquad d' = d^\beta + \Sigma x^\beta. \tag{77}$$

The example used to verify this conjugate relationship was constructed with a prior specified by hyperparameters $a = 2.0$, $b = 3.0$, $d = 2.5$ and the known scale parameter $\theta = 2.5$. The statistics describing the data are $n = 5$, and $P = 8.627$. In addition to these statistics, the data $1.54, 1.52, 1.53, 1.59, 1.515$ are needed to compute $\sum x^\beta$. Figure 16 shows the posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.12 Continuous Data: Lognormal Process

The lognormal process arises as a process transformation of the normal process. In order to directly illustrate the results discussed in section 1.2.1, there are no process reparameterizations employed. The probability density function for the lognormal process is

$$f(x \mid \mu, \rho) = \begin{cases} \frac{1}{y}\sqrt{\frac{\rho}{2\pi}} \exp\left(-\frac{\rho}{2}\left(\ln(x) - \mu\right)^2\right) & where \quad x > 0 \\ 0 & otherwise, \end{cases} \tag{78}$$

Note that parameters μ and ρ are not the mean and precision of the lognormal process.

2.12.1 Unknown Parameter μ

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a lognormal process where the parameter μ, is unknown and the parameter, ρ, is known. The likelihood function, $L(\mu \mid x_1, ..., x_n)$, proportional to μ is

$$L(\mu \mid x_1, ..., x_n) \propto \exp\left(-\frac{n\rho}{2}\left(\frac{\sum \ln(x)}{n} - \mu\right)^2\right). \tag{79}$$

The sufficient statistics are n, the number of data points, and $\overline{x} = \frac{\sum \ln(x)}{n}$, the mean of the log-data. As a function of μ, Equation 2.53 is proportional to a normal distribution of μ. The conjugate prior, $\pi(\cdot)$, is a normal distribution with hyperparameters m and $p > 0$

$$\pi(\mu \mid m, p) = \sqrt{\frac{p}{2\pi}} \exp\left(-\frac{p}{2}(\mu - m)^2\right). \tag{80}$$

The posterior distribution, $\pi(\mu \mid m', p')$, is a normal distribution with hyperparameters

$$m' = \frac{mp + n\rho\overline{x}}{p + n\rho} \qquad p' = p + n\rho. \tag{81}$$

The example used to verify this conjugate relationship was constructed with a normal prior specified by hyperparameters $m = 2.5$ and $p = 2.0$ with the known precision $\rho = 1.0$. The sufficient statistics describing the data are $n = 5$ and $\overline{x} = 1.856$. Figure 17 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.12.2 Unknown Parameter ρ

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a lognormal process where the parameter μ, is known and the parameter, ρ, is unknown. The likelihood function, $L(\rho \mid x_1, ..., x_n)$, proportional to ρ is

$$L(\rho \mid x_1, ..., x_n) \propto \rho^{\frac{n}{2}} \exp\left(-\frac{\rho}{2}\sum(\ln x - \mu)^2\right). \tag{82}$$

The sufficient statistics are n, the number of data points, and $SS = \sum(\ln x - \mu)^2$ the sum of the squared deviations of the log-data about μ. Viewed as a function ρ, Equation 2.56 is proportional to a gamma distribution. The conjugate prior distribution of the precision, $\pi(\cdot)$, is a gamma distribution with hyperparameters $\alpha > 0$ and $\beta > 0$

$$\pi(\rho \mid \alpha, \beta) = \begin{cases} \frac{\rho^{\alpha-1}\exp(\frac{-\rho}{\beta})}{\Gamma(\alpha)\beta^{\alpha}} & where \quad \rho > 0 \\ 0 & otherwise \end{cases}. \tag{83}$$

The posterior distribution, $\pi(\rho \mid \alpha', \beta')$, is a gamma distribution with hyperparameters

$$\alpha' = \alpha + \tfrac{n}{2} \qquad \beta' = \beta + \tfrac{SS}{2}. \tag{84}$$

The example used to verify this conjugate relationship was constructed with a gamma prior specified by hyperparameters $\alpha = 2.0$, $\beta = 1.5$, and the parameter μ =1.85. The sufficient statistics describing the data are $n = 5$ and $SS = 2.1046$. Figure 18 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.13 Continuous Directional Data: von Mises Process

Guttorp and Lockhart(1988) study conjugate prior relationships for univariate von Mises directional data. The von Mises process is specified by two parameters, its direction and concentration. The conjugate priors suggested for the two corresponding univariate priors are presented here.

2.13.1 Unknown Direction Parameter

Suppose that data $x_1, ..., x_n$ are independent and identically generated by a von Mises process where the concentration, $\kappa > 0$, is known and the direction, μ, is unknown. The likelihood function, $L(\mu \mid x_1, ..., x_n)$, proportional to μ is

$$L(\mu \mid x_1, ..., x_n) \propto \begin{cases} \exp\left(\kappa \sum_{i=1}^{n} \cos(x_i - \mu)\right) & where \quad 0 < x_i \leq 2\pi \\ 0 & otherwise \end{cases} . \tag{85}$$

The sufficient statistics are n, the number of data points, $\sum_{i=1}^{n} \cos(x_i)$, and $\sum_{i=1}^{n} \sin(x_i)$. The conjugate prior, $\pi(\cdot)$, is itself a von Mises distribution specified by concentration and direction hyperparameters, $a > 0$ and $0 < b \leq 2\pi$ respectively,

$$\pi(\mu \mid a, b) = \begin{cases} \frac{1}{I_0(a)} \exp\left(a \cos(\mu - b)\right) & where \quad 0 < \mu \leq 2\pi \\ 0 & otherwise \end{cases} , \tag{86}$$

where I_0 is a modified Bessel function of the first kind and of order 0. The posterior distribution of the unknown direction is equivalent to (2.60) with the updated hyperparameters

$$\begin{aligned} a' &= \kappa\left(a \sin(b) + \textstyle\sum_{i=1}^{n} \sin(x_i)\right) \\ b' &= \arctan\left(\frac{a \sin(b) + \sum_{i=1}^{n} \sin(x_i)}{a \cos(b) + \sum_{i=1}^{n} \cos(x_i)}\right). \end{aligned} \tag{87}$$

The example used to verify this conjugate relationship was constructed with a the prior specified by hyperparameters $a = 0.5$ and $b = 1.0472$ with the known concentration $\kappa = 1.5$. The sufficient statistics describing the data are $n = 5$, $\sum cos(x) = 1.1$, and $\sum sin(x) = 4.802$. Figure 19 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

2.13.2 Unknown Concentration Parameter

Suppose that data $x_1, ..., x_n$ are independent and identically generated by a von Mises process where the direction, $0 < \mu \leq 2\pi$, is known and the concentration, κ, is unknown. The likelihood function,

$L(\kappa \mid x_1, ...x_n)$, proportional to κ is

$$L(\kappa \mid x_1, ...x_n) \propto \begin{cases} \frac{1}{I_0(\kappa)^n} \exp\left(\kappa \sum_{i=1}^n \cos(x_i - \mu)\right) & where \quad 0 < x_i \leq 2\pi \\ 0 & otherwise \end{cases} . \tag{88}$$

where I_0 is a modified Bessel function of the first kind and of order 0. The sufficient statistics are n, the number of data points, and $\sum_{i=1}^n \cos(x_i - \mu)$. The conjugate prior, $\pi(\cdot)$, is specified by two hyperparameters, $c > 0$ and $0 < R_0 \leq c$,

$$\pi(\kappa \mid c, R_0) = \begin{cases} \frac{1}{K} \frac{\exp(\kappa R_0)}{I_0(\kappa)^c} & where \quad \kappa > 0 \\ 0 & otherwise \end{cases} , \tag{89}$$

$$where \qquad K = \int_0^\infty \frac{\exp(\kappa R_0)}{I_0(\kappa)^c} d\kappa.$$

R_0 can be thought of as the component on the x-axis of the resultant of c observations. The posterior distribution of the unknown mean is equivalent to (2.63) with the updated hyperparameters

$$c' = c + n \qquad R_0' = R_0 + \sum_{i=1}^n \cos(x_i - \mu). \tag{90}$$

The example used to verify this conjugate relationship was constructed with the prior specified by hyperparameters $c = 2.0$ and $R_0 = 1.5$ with the known direction $\mu = 0$. The sufficient statistics describing the data are $n = 5$ and $\sum_{i=1}^n \cos(x_i - \mu) = 4.983$. Figure 20 shows the conjugate posterior plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. The numerical posterior was generated with 100,000 samples from the prior distribution.

C:/WINAPP/SW20/DOC/CONJ;conjintr

chapter:3,page:1,section:3

3 Univariate Data & Bivariate Parameters

3.1 Continuous Data: Uniform Process

DeGroot(1970) discusses the conjugate prior for uniformly distributed data when both endpoints are unknown. Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a uniform process where both upper, u, and lower, l, boundaries are unknown. The likelihood function, $L(u, l \mid x_1, ...x_n)$, proportional to the boundary parameters, is

$$L(u, l|x_1, \ldots, x_n) \propto \begin{cases} \frac{1}{(u-l)^n} & where \quad l < \min(x_i), \max(x_i) < u \\ 0 & otherwise. \end{cases} \tag{91}$$

The sufficient statistics are n, the number of data points, $\max(x_i)$, the value of the maximum data point, and $\min(x_i)$, the value of the minimum data point. Viewed as a function of u & l, equation (3.1) is proportional to the bilateral bivariate Pareto distribution (DeGroot 1970). The conjugate prior, $\pi(\cdot,\cdot)$, is a bilateral bivariate Pareto distribution with hyperparameters $\alpha > 0$ and $r_1 > \min(x_i)$ and $r_2 < \max(x_i), r_1 < r_2$

$$\pi(u, l|r_1, r_2, \alpha) = \begin{cases} \frac{\alpha(\alpha+1)(r_2-r_1)^\alpha}{(u-l)^{\alpha+2}} & where \quad l < r_1, u > r_2 \\ 0 & otherwise. \end{cases} \tag{92}$$

If the joint distribution of u & l is given by (3.2) then the marginal distributions of $r_2 - l$ and $u - r_1$ are univariate Pareto distributions with parameters r_1, r_2, and α:

$$f(u - r_1|r_2 - r_1, \alpha) = \begin{cases} \frac{\alpha(r_2-r_1)^\alpha}{(u-r_1)^{\alpha+1}} & where \quad u > r_2 \\ 0 & otherwise \end{cases} \tag{93}$$

$$f(r_2 - l|r_2 - r_1, \alpha) = \begin{cases} \frac{\alpha(r_2-r_1)^\alpha}{(r_2-l)^{\alpha+1}} & where \quad l < r_1 \\ 0 & otherwise. \end{cases} \tag{94}$$

The posterior joint distribution of u & l is a bilateral bivariate Pareto distribution specified by the hyperparameters

$$\alpha' = \alpha + n \quad r_1' = \min(r_1, x_i) \quad r_2' = \max(r_2, x_i). \tag{95}$$

The example used to verify this conjugate relationship was constructed with a bilateral bivariate Pareto prior specified by hyperparameters $\alpha = 1.5$, $r_1 = -0.4$, and $r_2 = 0.1$. The sufficient statistics describing the data are $n = 5$, $\max(x) = 0.47$, and $\min(x) = -0.45$. Figure 21 shows the posterior marginal distribution of l plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. Figure 22 shows the corresponding posterior marginal distribution of u plotted with a 100 bin histogram. Each numerical joint posterior was generated with 600,000 samples from the prior distribution.

3.2 Continuous Data: Gamma Process

Miller (1980) developed a conjugate prior for the gamma data generating process. Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a gamma process where both the shape, α, and the reciprocal scale, β, parameters are unknown. The likelihood function, $L(\alpha, \beta \mid x_1, ...x_n)$, proportional to the parameters, is

$$L(\alpha, \beta|x_1, \ldots, x_n) \propto \begin{cases} \frac{P^{\alpha-1}\exp(-\beta S)}{(\Gamma(\alpha)\beta^{-\alpha})^n} & where \quad x_i > 0, i = 1 \ldots n \\ 0 & otherwise, \end{cases} \tag{96}$$

$$where \qquad S = \sum_{i=1}^{n} x_i \qquad P = \prod_{i=1}^{n} x_i.$$

The sufficient statistics are n, the number of data points, P, the product of the data, and S, the sum of the data. The factors of equation (3.6) proportional to parameters α & β make up the kernel of the conjugate prior, $\pi(\cdot,\cdot)$. We specify the conjugate prior with hyperparameters $p, q, r, s > 0$

$$\pi(\alpha, \beta|p, q, r, s) = \begin{cases} \frac{1}{K}\frac{p^{\alpha-1}\exp(-\beta q)}{\Gamma(\alpha)^r\beta^{-\alpha s}} & where \quad \alpha, \beta > 0 \\ 0 & otherwise. \end{cases} \tag{97}$$

The normalizing constant, K, is

$$K = \int_0^\infty \frac{p^{\alpha-1}\Gamma(s\alpha+1)}{\Gamma(\alpha)^r q^{\alpha s+1}} d\alpha. \tag{98}$$

An analytical form of the marginal distribution of the shape parameter, α, can be found,

$$f(\alpha|p,q,r,s) = \begin{cases} \frac{1}{K}\frac{p^{\alpha-1}\Gamma(s\alpha+1)}{q^{-\alpha s+1}\Gamma(\alpha)^r} & where \quad \alpha > 0 \\ 0 & otherwise, \end{cases} \tag{99}$$

with the normalizing constant, K, defined by (3.8). The marginal distribution of the scale parameter is found numerically by integrating α out of equation (3.7).

The posterior joint distribution of α and β is specified by the hyperparameters

$$p' = pP \quad q' = q + S \quad r' = r + n \quad s' = s + n. \tag{100}$$

The example used to verify this conjugate relationship was constructed with a prior specified by hyperparameters $p = 2.0$, $q = 1.230$, $r = 1.5$, and $s = 1.0$. The sufficient statistics describing the data are $n = 5$, $P = 1.3696$, and $S = 5.33$. Figure 23 shows the posterior marginal distribution of α plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. Figure 24 shows the corresponding posterior marginal distribution of β plotted with a 100 bin histogram. Each numerical joint posterior was generated with 600,000 samples from the prior distribution.

3.3 Continuous Data: Pareto Process

Arnold and Press(1983) discuss two priors for the Pareto data generating process where both the shape and precision parameters are unknown. The "Modified Lwin" prior is a conjugate prior distribution with dependence between the parameters. The other prior is developed so that one can specify prior independence of the unknown parameters. By enriching this "Independent" prior distribution by the prescriptions of Raiffa and Schlaiffer we are able achieve conjugacy. The censored data configurations considered by Arnold and Press are not discussed in this study.

3.3.1 Modified Lwin Prior

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Pareto process where both the shape, α, and the precision, τ, parameters are unknown. The likelihood function, $L(\alpha, \tau \mid x_1, ..., x_n)$, proportional to the parameters, is

$$L(\alpha, \tau|x_1, \ldots, x_n) \propto \begin{cases} \frac{\alpha^n}{\tau^{\alpha n} P^{\alpha+1}} & where \quad \min(x_i) > \frac{1}{\tau} \\ 0 & otherwise. \end{cases} \tag{101}$$

$$where \qquad P = \prod_{i=1}^{n} x_i.$$

The sufficient statistics are n, the number of data points and P, the product of the data. Viewed as a function of α & τ, equation (3.11) is proportional to the product of a Gamma distribution of α and a Pareto distribution of τ conditional on α. Given hyperparameters $a, c, d > 0$ and $b \leq \min(x_i)$ the conjugate prior, $\pi(\cdot, \cdot)$, is

$$\pi(\alpha, \tau | a, b, c, d) = \begin{cases} \frac{\alpha^{c-1} \exp(-\alpha d)}{\Gamma(c) d^{-c}} \frac{ab\alpha}{(b\tau)^{a\alpha+1}} & where \qquad \alpha > 0, \tau > \frac{1}{b} \\ 0 & otherwise. \end{cases} \tag{102}$$

The marginal distribution of the shape parameter, α, is simply

$$f(\alpha | c, d) = \begin{cases} \frac{\alpha^{c-1} \exp(-\alpha d)}{\Gamma(c) d^{-c}} & where \qquad \alpha > 0 \\ 0 & otherwise. \end{cases} \tag{103}$$

The marginal distribution of the precision parameter, τ, is

$$f(\tau | a, b, c, d) = \begin{cases} \frac{ac}{d\tau} \left(1 + \frac{a}{d} \ln(\tau b)\right)^{-c-1} & where \qquad \tau > \frac{1}{b} \\ 0 & otherwise. \end{cases} \tag{104}$$

The posterior joint distribution of α and τ is specified by the hyperparameters

$$\begin{array}{ll} a' = a + n & b' = b \\ c' = c + n & d' = d + \ln(P) - n\ln(b). \end{array} \tag{105}$$

The example used to verify this conjugate relationship was constructed with a prior specified by hyperparameters $a = 1.4168, b = 0.9$, $c = 2.0$, and $d = 2.17$. The sufficient statistics describing the data are $n = 10$ and $P = 6.0378$. Figure 25 shows the posterior marginal distribution of α plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. Figure 26 shows the corresponding posterior marginal distribution of τ plotted with a 100 bin histogram. Each numerical joint posterior was generated with 600,000 samples from the prior distribution.

3.3.2 Independent Prior

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Pareto process where both the shape, α, and the precision, τ, parameters are unknown. The likelihood function, $L(\alpha, \tau \mid x_1, ..., x_n)$, proportional to the parameters, is

$$L(\alpha, \tau | x_1, \ldots, x_n) \propto \begin{cases} \frac{\alpha^n}{\tau^{\alpha n} P^{\alpha+1}} & where \qquad \min(x_i) > \frac{1}{\tau} \\ 0 & otherwise, \end{cases} \tag{106}$$

$$where \qquad P = \prod_{i=1}^{n} x_i.$$

The sufficient statistics are n, the number of data points and P, the product of the data. Since the likelihood function is determined by the process, it remains the same as (3.11). Therefore, the conjugate prior must still be able to adopt a form proportional to (3.11). This prior is constructed by adding another hyperparameter which, depending on its value, allows τ and α to be uncoupled.

Given hyperparameters a, $c, d > 0, b \leq \min(x_i)$, and $m \geq 0$ the "Independent" conjugate prior, $\pi(\cdot,\cdot)$, is

$$\pi(\alpha,\tau|a,b,c,d,m) = \begin{cases} \frac{1}{K}\frac{\alpha^{c-1}\exp(-\alpha d)}{\tau^{\alpha m+a+1}} & where \quad \alpha > 0, \tau > \frac{1}{b} \\ 0 & otherwise. \end{cases} \tag{107}$$

The normalizing constant, K, is

$$K = \int_{\frac{1}{b}}^{\infty}\int_{0}^{\infty} \frac{\alpha^{c-1}\exp(-\alpha d)}{\tau^{\alpha m+a+1}} d\alpha\, d\tau. \tag{108}$$

The marginal distribution of the shape parameter, α, is

$$f(\alpha|a,b,c,d,m) = \begin{cases} \frac{1}{K}\frac{\alpha^{c-1}\exp(-\alpha(d-m\ln(b)))}{a+m\alpha} & where \quad \alpha > 0 \\ 0 & otherwise, \end{cases} \tag{109}$$

where K, the normalizing constant, is defined by equation (3.18). The marginal distribution of the precision parameter, τ, is

$$f(\tau|a,b,c,d,m) = \begin{cases} \frac{1}{K}\frac{\Gamma(c)}{\tau^{a+1}(d+m\ln(\tau))^{c}} & where \quad \tau > \frac{1}{b} \\ 0 & otherwise. \end{cases} \tag{110}$$

where K, the normalizing constant, is defined by equation (3.18).

The posterior joint distribution of α and τ is specified by the hyperparameters

$$\begin{array}{ll} a' = a+n & b' = \min(b, \min(x_i)) \\ c' = c+n & d' = d+\ln(P) \\ m' = m+n. & \end{array} \tag{111}$$

The example used to verify this conjugate relationship was constructed with a prior specified by hyperparameters $a = 1.4168, b = 0.9, c = 2.0, d = 2.17$, and $m = 0$. Note that there is prior independence between the shape and precision parameters because m equals zero. The sufficient statistics describing the data are $n = 10$ and $P = 6.0378$. Figure 27 shows the posterior marginal distribution of α plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. Figure 28 shows the corresponding posterior marginal distribution of τ plotted with a 100 bin histogram. Each numerical joint posterior was generated with 600,000 samples from the prior distribution.

3.4 Continuous Data: Normal Process

The Gamma-Normal distribution is the most common conjugate prior for the simultaneous inference of mean and precision parameters of a Normal process. We also present two extensions of the Gamma-Normal that have additional convenient properties. What we call Athreya's Extended Gamma is an extension of the Gamma marginal on the precision (Athreya 1988). This extension allows one to parameterize the process variability in terms of either the variance or the precision, and to easily switch between the two. Dickey's(1980) Gamma-Normal extension allows one to specify prior independence of the mean and precision parameters. Although such a prior is not dealt with here, one could combine these two extensions to create a conjugate prior with all of the aforementioned properties. We begin by describing the unmodified Gamma-Normal prior.

3.4.1 Gamma-Normal

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Normal process where both the mean, m, and the precision, p, parameters are unknown. The likelihood function, $L(m, p \mid x_1, ..., x_n)$, proportional to the parameters, is

$$L(m, p \mid x_1, ..., x_n) \propto p^{\frac{n}{2}} \exp\left(-\frac{p}{2}\sum_{i=1}^{n}(x_i - m)^2\right). \tag{112}$$

The sufficient statistics are n, the number of data points, $\overline{x}$, the mean of the data, and SS, the sum of the squared deviations of the data. Viewed as a function of m & p, Equation 3.22 is proportional to the product of a Gamma distribution of p and a Normal distribution of m conditional on p. Given hyperparameters α,β,$\tau > 0$ and μ the conjugate prior, $\pi(\cdot, \cdot)$, is

$$\pi(m, p \mid \alpha, \beta, \tau, \mu) = \begin{cases} \frac{p^{\alpha-1}\exp\left(-\frac{p}{\beta}\right)}{\Gamma(\alpha)\beta^{\alpha}} \times & \\ \left(\frac{p\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{p\tau}{2}(m-\mu)^2\right) & where \quad p > 0 \\ 0 & otherwise. \end{cases} \tag{113}$$

The marginal distribution of the precision parameter, p, is simply

$$f(p \mid \alpha, \beta) = \begin{cases} \frac{p^{\alpha-1}\exp\left(-\frac{p}{\beta}\right)}{\Gamma(\alpha)\beta^{\alpha}} & where \quad p > 0 \\ 0 & otherwise \end{cases}. \tag{114}$$

The marginal distribution of the mean parameter, m, is a t distribution with 2α degrees of freedom, location μ, and precision $\alpha\tau\beta$:

$$f(m \mid \alpha, \beta, \tau, \mu) = \sqrt{\frac{\beta\tau}{2\pi}}\frac{\Gamma\left(\frac{2\alpha+1}{2}\right)}{\Gamma(\alpha)}\left(1 + \frac{\tau\beta}{2}(m-\mu)^2\right)^{-\frac{2\alpha+1}{2}}. \tag{115}$$

The posterior joint distribution of m and p is specified by the hyperparameters

$$\alpha' = \alpha + \frac{n}{2} \qquad \beta' = \left(\frac{1}{\beta} + \frac{SS}{2} + \frac{\tau n(\overline{x}-\mu)^2}{2(\tau+n)}\right)^{-1}$$

$$\mu' = \frac{\tau\mu + n\overline{x}}{\tau + n} \qquad \tau' = \tau + n. \tag{116}$$

The example used to verify this conjugate relationship was constructed with a prior specified by hyperparameters $\alpha = 3.0$, $\beta = 1.0$, $\tau = 0.1$,and $\mu = 2.0$. The sufficient statistics describing the data are $n = 10, SS = 5.4$, and $\overline{x} = 4.2$. Figure 29 shows the posterior marginal distribution of m plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. Figure 30 shows the corresponding posterior marginal distribution of p plotted with a 100 bin histogram. Each numerical joint posterior was generated with 600,000 samples from the prior distribution.

3.4.2 Athreya's Extended Gamma

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Normal process where both the mean, m, and the variance, v, parameters are unknown. The likelihood function, $L(m, v \mid x_1, ..., x_n)$, proportional to the parameters, is

$$L(m, v \mid x_1, ..., x_n) \propto v^{-\frac{n}{2}} \exp\left(-\frac{1}{2v}\sum_{i=1}^{n}(x_i - m)^2\right). \tag{117}$$

The sufficient statistics are n, the number of data points, $\overline{x}$, the mean of the data, and SS, the sum of the squared deviations of the data. Like the Gamma-Normal distribution Athreya's prior is product of a marginal distribution of the variance and a conditional distribution of the mean. The marginal distribution of the variance is a three parameter Gamma-like distribution known as the generalized inverse Gaussian distribution. The conditional distribution of the mean is a Normal distribution. Given hyperparameters $\alpha, \beta, \gamma, \tau > 0$ and μ the conjugate prior, $\pi(\cdot, \cdot)$, is

$$\pi(m, v \mid \alpha, \beta, \gamma, \tau, \mu) = \begin{cases} \frac{1}{K} v^{\alpha} \exp\left(-\frac{\beta}{v} - v\gamma\right) \times & \\ (v)^{-\frac{1}{2}} \exp\left(-\frac{1}{2v\tau}(m-\mu)^2\right) & where \qquad v > 0 \\ 0 & otherwise \end{cases} \tag{118}$$

We determined an analytic form of the normalizing constant where both β and γ are greater than zero (Gradshteyn & Ryzhik section 3.471 equation 9)

$$K = \left(2\sqrt{2\pi\tau}\left(\frac{\beta}{\gamma}\right)^{\frac{\alpha+1}{2}} K_{\alpha+1}\left(2\sqrt{\beta\gamma}\right)\right)^{-1}, \tag{119}$$

where K_α is a modified Bessel function of the second kind and of order α. If either hyperparameter β or γ is zero the normalizing constant simplifies.

The marginal distribution of the variance parameter, v, is

$$f(v \mid \alpha, \beta, \gamma, \tau) = \begin{cases} \frac{\sqrt{2\pi\tau}}{K} v^{\alpha} \exp\left(-\frac{\beta}{v} - v\gamma\right) & where \qquad v > 0 \\ 0 & otherwise \end{cases}, \tag{120}$$

with the normalizing constant, K , defined by (3.29).

Integrating the variance parameter out of (3.28) yields the marginal distribution of the mean. We determined the analytical result of this integration, a modified Bessel function of the second kind (Gradshteyn & Ryzhik section 3.471 equation 9),

$$f(m \mid \alpha, \beta, \gamma, \tau, \mu) = \frac{\frac{2}{K}\left(\frac{\beta + \frac{1}{2\tau}(m-\mu)^2}{\gamma}\right)^{\frac{\alpha+.5}{2}} \times}{K_{\alpha+.5}\left(2\sqrt{\gamma\left(\beta + \frac{1}{2\tau}(m-\mu)^2\right)}\right)}, \tag{121}$$

with the normalizing constant, K,defined by (3.29).

The posterior joint distribution of m and v is specified by the hyperparameters

$$\alpha' = \alpha - \frac{n}{2} \qquad \beta' = \frac{1}{\beta} + \frac{SS}{2} + \frac{n\left(\overline{x} - \mu\right)^2}{2\left(1 + n\tau\right)}$$

$$\mu' = \frac{\mu + \tau n \overline{x}}{1 + n\tau} \qquad \tau' = \frac{\tau}{1 + n\tau} \qquad \gamma' = \gamma. \tag{122}$$

The example used to verify this conjugate relationship was constructed with a prior specified by hyperparameters $\alpha = 8.0$, $\beta = 2.1$, $\gamma = 1.0$, $\tau = 2.0$,and $\mu = 2.0$. The sufficient statistics describing the data are $n = 10$, $SS = 5.4$, and $\overline{x} = 4.2$. Figure 31 shows the posterior marginal distribution of m plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. Figure 32 shows the corresponding posterior marginal distribution of v plotted with a 100 bin histogram. Each numerical joint posterior was generated with 600,000 samples from the prior distribution.

3.4.3 Dickey's Gamma-Normal

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Normal process where both the mean, m, and the precision, p, parameters are unknown. The likelihood function, $L(m, p \mid x_1, ..., x_n)$, proportional to the parameters, is

$$L(m, p \mid x_1, ..., x_n) \propto p^{\frac{n}{2}} \exp\left(-\frac{p}{2}\sum_{i=1}^{n}\left(x_i - m\right)^2\right). \tag{123}$$

The sufficient statistics are n, the number of data points, $\overline{x}$, the mean of the data, and SS, the sum of the squared deviations of the data. Dickey prescribes a conjugate prior that is the product of the Gamma-Normal (3.22) and a Normal distribution of the unknown mean. Given the Gamma-Normal hyperparameters α, $\beta > 0$, $\tau \geq 0$, and μ_1, and the two additional hyperparameters for the Normal, $\gamma > 0$ and μ_2 the conjugate prior, $\pi\left(\cdot, \cdot\right)$, is

$$\pi(m, p \mid \alpha, \beta, \tau, \mu_1, \gamma, \mu_2) = \begin{cases} \frac{p^{\alpha-1}\exp\left(-\frac{p}{\beta}\right)}{\Gamma(\alpha)\beta^{\alpha}}\left(\frac{p\tau}{2\pi}\right)^{\frac{\Delta(\tau)}{2}} \times & \\ \exp\left(-\frac{p\tau}{2}\left(m - \mu_1\right)^2\right) \times & \\ \sqrt{\frac{\gamma}{2\pi}}\exp\left(-\frac{\gamma}{2}\left(m - \mu_2\right)^2\right) & where \qquad p > 0 \\ 0 & otherwise, \end{cases} \tag{124}$$

where Δ is the Delta function. Notice that if $\tau = 0$, (3.34) reduces to the product of independent Gamma and Normal marginals of p and m, respectively.

The marginal distribution of the precision parameter, p, is

$$f(p \mid \alpha, \beta, \tau, \mu_1, \gamma, \mu_2) = \begin{cases} \frac{1}{K}p^{\alpha + \frac{\Delta(\tau)}{2} - 1}\left(p\tau + \gamma\right)^{-\frac{\Delta(\tau)}{2}} \times & \\ \left(\frac{p\tau}{2\pi}\right)\exp\left(-\frac{p}{\beta}\right) \times & \\ \exp\left(-\frac{p\tau\gamma}{2(p\tau+\gamma)}\left(\mu_1 - \mu_2\right)^2\right) & where \qquad p > 0 \\ 0 & otherwise, \end{cases} \tag{125}$$

where the normalizing constant, K, is equal to (3.35) integrated on variable p over the range zero to infinity.

The marginal distribution of the mean parameter, m, is

$$f(m \mid \alpha, \beta, \tau, \mu_1, \gamma, \mu_2) = \frac{1}{K}\left(\frac{1}{\beta} + \frac{\tau}{2}(m-\mu_1)^2\right)^{-\frac{\Delta(\tau)}{2}(2\alpha+1)} \exp\left(-\frac{\gamma}{2}(m-\mu_2)^2\right), \tag{126}$$

where the normalizing constant, K, is equal (3.36) integrated on variable m over the range negative infinity to positive infinity.

The posterior joint distribution of m and p is specified by the hyperparameters

$$\alpha' = \alpha + \frac{n}{2} \qquad \beta' = \left(\frac{1}{\beta} + \frac{SS}{2} + \frac{\tau n(\overline{x}-\mu_1)^2}{2(\tau+n)}\right)^{-1}$$

$$\mu_1' = \frac{\tau\mu_1 + n\overline{x}}{\tau+n} \qquad \tau' = \tau + n \tag{127}$$

$$\mu_2' = \mu_2 \qquad \gamma' = \gamma.$$

The example used to verify this conjugate relationship was constructed with a prior specified by hyperparameters $\alpha = 2.0$, $\beta = 2.0$, $\tau = 1.0$, $\mu_1 = 3.0$, $\mu_2 = 2.0$, and $\gamma = 0.50$. The sufficient statistics describing the data are $n = 5$, $SS = 0.362$, and $\overline{x} = 1.39$. Figure 33 shows the posterior marginal distribution of m plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. Figure 34 shows the corresponding posterior marginal distribution of p plotted with a 100 bin histogram. Each numerical joint posterior was generated with 600,000 samples from the prior distribution.

3.5 Continuous Data: Inverse Gaussian Process

Banerjee and Bhattacharyya(1979) discuss a conjugate prior for the Inverse Gaussian Process with both parameters unknown. Although these authors noted the similarity between their conjugate prior and the Gamma-Normal distribution they nowhere note that their prior *is* a Gamma-Normal. The Gamma-Normal prior is used here.

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Inverse Gaussian process where both the reciprocal mean parameter, m, and the dispersion parameter, λ, are unknown. The likelihood function, $L(m, \lambda \mid x_1, ..., x_n)$, proportional to the parameters, is

$$L(m, \lambda \mid x_1, ..., x_n) \propto \lambda^{\frac{n}{2}} \exp\left(-\frac{\lambda m^2}{2}\left(\sum x - \frac{2n}{m} + \frac{\sum \frac{1}{x}}{m^2}\right)^2\right). \tag{128}$$

The sufficient statistics are n, the number of data points, $\sum x$, the sum of the data, and $\sum \frac{1}{x}$, the sum of the reciprocal data. Viewed as a function of m & λ, Equation 3.38 is proportional to the product of a Gamma distribution of λ and a Normal distribution of m conditional on λ. Given hyperparameters α,β,$\tau > 0$ and μ the conjugate prior, $\pi(\cdot, \cdot)$, is

$$\pi(m, \lambda \mid \alpha, \beta, \tau, \mu) = \begin{cases} \frac{\lambda^{\alpha-1}\exp(-\lambda\beta)}{\Gamma(\alpha)\beta^{-\alpha}} \times & \\ \left(\frac{\lambda\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\lambda\tau}{2}(m-\mu)^2\right) & where \quad \lambda > 0 \\ 0 & otherwise. \end{cases} \tag{129}$$

The marginal distribution of parameter, λ, is simply

$$f(\lambda \mid \alpha, \beta) = \begin{cases} \frac{\lambda^{\alpha-1}\exp(-\lambda\beta)}{\Gamma(\alpha)\beta^{-\alpha}} & where \quad \lambda > 0 \\ 0 & otherwise \end{cases}. \tag{130}$$

The marginal distribution of the reciprocal mean parameter, m, is a t distribution with 2α degrees of freedom, location μ, and precision $\frac{\alpha\tau}{\beta}$:

$$f(m \mid \alpha, \beta, \tau, \mu) = \sqrt{\frac{\tau}{2\pi\beta}} \frac{\Gamma\left(\frac{2\alpha+1}{2}\right)}{\Gamma(\alpha)} \left(1 + \frac{\tau}{2\beta}(m-\mu)^2\right)^{-\frac{2\alpha+1}{2}}. \tag{131}$$

The posterior joint distribution of m and p is specified by the hyperparameters

$$\alpha' = \alpha + \frac{n}{2} \qquad \beta' = \beta + \frac{\mu^2\tau + \sum \frac{1}{x}}{2} - \frac{(\mu\tau + n)^2}{2(\tau + \sum x)}$$

$$\mu' = \frac{\tau\mu + n}{\tau + \sum x} \qquad \tau' = \tau + \sum x. \tag{132}$$

The example used to verify this conjugate relationship was constructed with a prior specified by hyperparameters $\alpha = 2.0$, $\beta = 0.33$, $\tau = 1.0$,and $\mu = 4.0$. The sufficient statistics describing the data are $n = 5, \sum x = 2.695$, and $\sum \frac{1}{x} = 9.298$. Figure 35 shows the posterior marginal distribution of m plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. Figure 36 shows the corresponding posterior marginal distribution of p plotted with a 100 bin histogram. Each numerical joint posterior was generated with 600,000 samples from the prior distribution.

3.6 Continuous Data: Weibull Process

Although the prior presented here is not a true conjugate prior, it's posterior kernel is analytically tractable and it's normalizing constant is easily computed with standard numerical integration routines. The posterior distribution is seen to have the same general functional form as the prior.

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a Weibull process where both the shape, β, and scale, θ, parameters are unknown. The likelihood function, $L(\beta, \theta \mid x_1, ..., x_n)$, proportional to the parameters, is

$$L(\beta, \theta \mid x_1, ..., x_n) \propto \left(\frac{\beta}{\theta}\right)^n \left(\prod_{i=1}^{n} x_i\right)^{\beta-1} \exp\left(-\frac{\sum x^\beta}{\theta}\right). \tag{133}$$

The statistics of interest are n, the number of data points, $\sum x^{\beta}$, the sum of the data raised to the βth power, and $\prod x$, the product of the data. Viewed as a function of β & θ, Equation 3.43 is conformable with the following prior distribution, $\pi\left(\beta, \theta \mid a, b, c, \mathbf{d}, v\right)$, where the hyperparameters a,b,$c > 0$ and ν is a non-negative integer indicator. Special consideration of the hyperparameters $\mathbf{d}$ & v must be made because they, along with function $D(\beta, \nu, \mathbf{d})$, allow us to deal with the lack of a set of fixed dimension sufficient statistics. Let $\mathbf{d} = (d_0, x_1, ..., x_n)^T$, ν be the number of these elements that we will consider, and let the function $D\left(\cdot\right)$ be

$$D(\beta, \nu, \mathbf{d}) = \begin{cases} d_0^{\beta} & where \quad \nu = 0 \\ d_0^{\beta} + \sum_{i=1}^{\nu} x_i^{\beta} & where \quad \nu = 1, 2, ... \end{cases} \tag{134}$$

Thus, the prior distribution is specified by the hyperparameter values a, b, c, d_0, and $\nu = 0$:

$$\pi(\beta, \theta \mid a, b, c, \mathbf{d}, v) = \begin{cases} \frac{1}{K}\beta^{a-1}\exp\left(-\beta b\right)\theta^{-c}\exp\left(-\frac{D(\beta,\nu,\mathbf{d})}{\theta}\right) & where \quad \beta, \theta > 0 \\ 0 & otherwise. \end{cases} \tag{135}$$

The normalizing constant, K, is

$$K = \int_0^{\infty}\int_0^{\infty} \beta^{a-1}\exp\left(-\beta b\right)\theta^{-c}\exp\left(-\frac{D(\beta, \nu, \mathbf{d})}{\theta}\right)\, d\theta\; d\beta. \tag{136}$$

The marginal distribution of the shape parameter, β, is

$$f(\beta \mid a, b, c, d, \nu) = \begin{cases} \frac{\Gamma(c-1)}{K}\beta^{a-1}\exp\left(-\beta b\right)\; D(\beta, \nu, \mathbf{d}) & where \quad \beta > 0 \\ 0 & otherwise, \end{cases} \tag{137}$$

where the normalizing constant, K, is defined by (3.46). The marginal distribution of the scale parameter, θ, is found numerically by integrating β out of Equation 3.45.

The posterior joint distribution of β and θ is specified by the hyperparameters

$$a' = a + n \qquad b' = b - \ln\left(\prod x\right) \qquad c' = c + n \qquad d_0' = d_0 \qquad \nu' = n \tag{138}$$

The example used to verify this conjugate relationship was constructed with a prior specified by hyperparameters $a = 20.0$, $b = 2.0$, $c = 6.0$, $\mathbf{d} = (2.5, 0.9, 1.52, 1.10)$, and $\nu = 0$. The statistics of interest describing the data are $n = 3$ and $\prod x = 1.5048$. Figure 37 shows the posterior marginal distribution of β plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. Figure 38 shows the corresponding posterior marginal distribution of θ plotted with a 100 bin histogram. Each numerical joint posterior was generated with 600,000 samples from the prior distribution.

3.7 Continuous Data: Lognormal Process

This conjugate prior relationship presented here illustrated the results presented in section 1.2.1 concerning conjugate priors under a process transformation. Since the lognormal process is a process transformation of the normal process, the joint conjugate prior of the lognormal process is a Gamma-Normal distribution.

Suppose that data $x_1, ..., x_n$ are independent and identically distributed from a logormal process where both parameters m and p are unknown. The likelihood function, $L(m, p \mid x_1, ..., x_n)$, proportional to the parameters, is

$$L(m, p \mid x_1, ..., x_n) \propto p^{\frac{n}{2}} \exp\left(-\frac{p}{2}\sum_{i=1}^{n}(\ln(x_i) - m)^2\right). \tag{139}$$

The sufficient statistics are n, the number of data points, $\overline{x} = \frac{\sum \ln(x)}{n}$, the mean of the log-data, and SS, the sum of the squared deviations of the log-data about m. Viewed as a function of m & p, Equation 3.49 is proportional to the product of a Gamma distribution of p and a Normal distribution of m conditional on p. Given hyperparameters α,β,$\tau > 0$ and μ the conjugate prior, $\pi(\cdot, \cdot)$, is

$$\pi(m, p \mid \alpha, \beta, \tau, \mu) = \begin{cases} \frac{p^{\alpha-1}\exp\left(-\frac{p}{\beta}\right)}{\Gamma(\alpha)\beta^{\alpha}} \times & \\ \left(\frac{p\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{p\tau}{2}(m-\mu)^2\right) & where \quad p > 0 \\ 0 & otherwise. \end{cases} \tag{140}$$

The marginal distribution of the precision parameter, p, is simply

$$f(p \mid \alpha, \beta) = \begin{cases} \frac{p^{\alpha-1}\exp\left(-\frac{p}{\beta}\right)}{\Gamma(\alpha)\beta^{\alpha}} & where \quad p > 0 \\ 0 & otherwise \end{cases}. \tag{141}$$

The marginal distribution of the mean parameter, m, is a t distribution with 2α degrees of freedom, location μ, and precision $\alpha\tau\beta$:

$$f(m \mid \alpha, \beta, \tau, \mu) = \sqrt{\frac{\beta\tau}{2\pi}} \frac{\Gamma\left(\frac{2\alpha+1}{2}\right)}{\Gamma(\alpha)} \left(1 + \frac{\tau\beta}{2}(m-\mu)^2\right)^{-\frac{2\alpha+1}{2}}. \tag{142}$$

The posterior joint distribution of m and p is specified by the hyperparameters

$$\alpha' = \alpha + \frac{n}{2} \qquad \beta' = \left(\frac{1}{\beta} + \frac{SS}{2} + \frac{\tau n(\overline{x}-\mu)^2}{2(\tau+n)}\right)^{-1}$$

$$\mu' = \frac{\tau\mu + n\overline{x}}{\tau + n} \qquad \tau' = \tau + n. \tag{143}$$

The example used to verify this conjugate relationship was constructed with a prior specified by hyperparameters $\alpha = 3.0$, $\beta = 1.0$, $\tau = 0.1$,and $\mu = 2.0$. The sufficient statistics describing the data are $n = 10, SS = 5.4$, and $\overline{x} = 4.2$. Figure 39 shows the posterior marginal distribution of m plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. Figure 40 shows the corresponding posterior marginal distribution of p plotted with a 100 bin histogram. Each numerical joint posterior was generated with 600,000 samples from the prior distribution.

3.8 Continuous Directional Data: von Mises Process

Guttorp and Lockhart(1988) developed a bivariate conjugate prior for the von Mises distribution with unknown concentration and direction. Mistakes in their normalizing constants and the marginal distributions have been corrected.

Suppose that data $x_1, ..., x_n$ are independent and identically generated by a von Mises process where both the concentration, κ, and the direction, μ, are unknown. The likelihood function, $L(\mu, \kappa \mid x_1, ..., x_n)$, proportional to κ & μ is

$$L(\mu, \kappa \mid x_1, ..., x_n) \propto \begin{cases} \frac{1}{I_0(\kappa)^n} \exp\left(\kappa \sum_{i=1}^{n} \cos(x_i - \mu)\right) & where \quad 0 < x_i \le 2\pi \\ 0 & otherwise. \end{cases} \tag{144}$$

The sufficient statistics are n, the number of data points, $\sum_{i=1}^{n} cos(x_i)$, and $\sum_{i=1}^{n} sin(x_i)$. The factors of Equation 3.54 proportional to parameters μ & κ make up the kernel of the conjugate prior, $\pi(\cdot, \cdot)$. We define the conjugate prior with hyperparameters R_0, $c > 0$, and $0 < \phi \le 2\pi$ as

$$\pi(\mu, \kappa \mid R_0, c, \phi) = \begin{cases} \frac{1}{K} \frac{\exp(\kappa R_0 \cos(\phi - \mu))}{I_0(\kappa)^c} & where \quad 0 < \mu \le 2\pi, \kappa > 0 \\ 0 & otherwise \end{cases}. \tag{145}$$

The normalizing constant K is

$$K = 2\pi \int_0^{\infty} \frac{I_0(R_0\kappa)}{I_0(\kappa)^c} d\kappa, \tag{146}$$

where I_0 is a modified Bessel function of the first kind and of order 0.

An analytical form of the marginal distribution of the concentration parameter, κ, was found,

$$f(\kappa \mid R_0, c) = \begin{cases} \frac{2\pi}{K} \frac{I_0(R_0\kappa)}{I_0(\kappa)^c} & where \quad \kappa > 0 \\ 0 & otherwise \end{cases}, \tag{147}$$

with the normalizing constant defined by (3.56). The marginal distribution of the direction parameter is found numerically by integrating κ out of equation (3.55).

The posterior distribution of the unknown direction and concentration parameters is equivalent to (3.55) with the updated hyperparameters

$$R_0' = \sqrt{\left(R_0 \sin(\phi) + \sum_{i=1}^{n} \sin(x_i)\right)^2 + \left(R_0 \cos(\phi) + \sum_{i=1}^{n} \cos(x_i)\right)^2}$$

$$\phi' = \arctan\left(\frac{R_0 \sin(\phi) + \sum_{i=1}^{n} \sin(x_i)}{R_0 \cos(\phi) + \sum_{i=1}^{n} \cos(x_i)}\right) \tag{148}$$

$$c' = c + n.$$

The example used to verify this conjugate relationship was constructed with a prior specified by hyperparameters $c = 2.0$, $R_0 = 1.25$, and $\phi = \pi/3$. The sufficient statistics describing the data are $n = 5$, $\sum cos(x) = 2.4154$, and $\sum sin(x) = 4.3712$. Figure 41 shows the posterior marginal distribution of μ plotted with a 100 bin histogram produced from the numerical estimate of the posterior distribution. Figure 42 shows the corresponding posterior marginal distribution of κ plotted with a 100 bin histogram. Each numerical joint posterior was generated with 600,000 samples from the prior distribution.

4 Summary of Conjugate Relationship Verifications

The similarity between each pair of numerically approximated posterior histograms proves that the numerical approximation is stable for the number of samples used with 100 bins. Furthermore, the similarity between the histograms and their associated true conjugate posterior distributions demonstrates that the numerical approximations are stably converging to the true conjugate posteriors. This verifies the conjugate relationships and their associated updating formulas. This verification also gives us a general feel for the difference in computational costs between the conjugate priors and numerical methods. The conjugate prior calculations require only the calculation of the updating formulas to produce the true posterior given the prior. For the same situation, the numerical method used requires computing a random deviate from the prior, calculating the likelihood of this sampled prior, and then binning this weight - several thousand times - to produce an approximation of the true posterior.

5 Multivariate Data & Multivariate Parameters

5.1 Discrete Data: Multinomial Process

DeGroot(1970) discusses the conjugate prior for data with a multinomial distribution. Suppose that data vector $\mathbf{x} = (x_1, ..., x_k)'$ is independent and identically distributed from a multinomial process where the parameters $\mathbf{w} = (w_1, ..., w_k)'$ are unknown. The likelihood function, $L(\mathbf{w} \mid \mathbf{x})$, proportional to the unknown parameters, is

$$L(\mathbf{w} \mid \mathbf{x}) \propto \begin{cases} \prod_{i=1}^{k} w_i^{x_i} & where \quad w_i > 0 \quad and \quad \sum_{i=1}^{k} w_i = 1 \\ 0 & otherwise. \end{cases} \tag{149}$$

The sufficient statistics are the x_i's. Viewed as a function of $\mathbf{w}$, Equation 4.1 is proportional to the Dirichlet distribution. The conjugate prior, $\pi(\cdot)$, is a Dirichlet distribution with hyperparameters $\alpha = (\alpha_1, ..., \alpha_k)'$ where $\alpha_i > 0$ and $\sum_{i=1}^{k} \alpha_i = 1$,

$$\pi(\mathbf{w} \mid \alpha) = \begin{cases} \frac{\Gamma(\alpha_1+...+\alpha_k)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} w_1^{\alpha_1 - 1} \cdots w_k^{\alpha_k - 1} & where \quad w_i > 0 \quad and \quad \sum_{i=1}^{k} w_i = 1 \\ 0 & otherwise. \end{cases} \tag{150}$$

If the joint distribution $\mathbf{w}$ is given by (4.2) then the marginal distributions of any w_j are Beta distributions with parameters α_j and $\sum_{i \neq j} \alpha_i - \alpha_j$. This result is generalized in DeGroot (1970 p. 50) to situations where $\pi(\mathbf{w} \mid \alpha)$ is marginalized such that it has between 2 and k categories. (Expand on this result?) The posterior joint distribution of $\mathbf{w}$ is a Dirichlet distribution specified by the hyperparameters

$$\alpha_i^* = \alpha_i + x_i \quad for \quad i = 1...k \tag{151}$$

Note that this result is a direct extension of the conjugate relationship for the Binomial process.

5.2 Discrete Data: Multivariate Hypergeometric Process

Suppose that data $\mathbf{x} = (x_1, ..., x_n)'$ arise from a multivariate hypergeometric process where process parameter N, the total population size, is known $(N > 0)$ and $\mathbf{M} = (M_1, ..., M_k)$, the vector of members in the k mutually exclusive and exhastive categories that make up the population, are unknown $(0 \leq M_i \leq N \quad and \quad \sum M_i = N)$. If $n = \sum x_i$ objects are sampled without replacement, x_i is the number of objects from category k. The likelihood function, $L(\mathbf{M} \mid \mathbf{x})$, proportional to the parameters, is

$$L(\mathbf{M} \mid \mathbf{x}) \propto \begin{cases} M_1 x_1 \cdots M_k x_k & where \quad 0 \leq x_i \leq \min(M_i, n), i = 1 \ldots k \\ 0 & otherwise. \end{cases} \tag{152}$$

Equation 4.4 is proportional to the Dirichlet-Multinomical distribution of $\mathbf{M}$. Consequently, the conjugate prior, $\pi(\cdot)$, is a Dirichlet-Multinomial distribution with hyperparameters $\alpha = (\alpha_1, ..., \alpha_k)'$ where $\alpha_i > 0$ and $\sum_{i=1}^{k} \alpha_i = 1$ and H, an integer greater than zero:

$$\begin{aligned} \pi(\mathbf{M} \quad | \quad \alpha{,}n) &= \frac{n!}{M_1! M_2! \cdots M_3!} \frac{\Gamma\left(\sum \alpha_i\right)}{\prod \Gamma\left(\alpha_i\right)} \frac{\prod \Gamma\left(\alpha_i + M_i\right)}{\Gamma\left(\sum\left(\alpha_i + M_i\right)\right)} \\ where \quad M_i \quad > \quad 0 \quad and \quad & \sum_{i=1}^{k} M_i = n \end{aligned} \tag{153}$$

The closure property is verified by showing that the posterior distribution is also a member of the Dirichlet-Multinomial family. The posterior distribution is the Dirichlet-Multinomial distribution:

$$\begin{aligned} \pi(\mathbf{M} \quad | \quad \mathbf{x}, \alpha{,}H) &= \frac{(N-n)!}{(M_1 - x_1)! \cdots (M_k - x_k)!} \frac{\Gamma\left(n + \sum \alpha_i\right)}{\prod \Gamma\left(\alpha_i + x_i\right)} \frac{\prod \Gamma\left(\alpha_i + M_i\right)}{\Gamma\left(\sum \alpha_i + N\right)} \\ where \quad M_i \quad > \quad 0 \quad and \quad & \sum_{i=1}^{k} M_i = H \end{aligned} \tag{154}$$

Note that this result is a direct extension of the conjugate relationship for the Binomial process.

5.3 Continuous Data: Multivariate Normal Process

I will begin this section by transcribing the theorems from DeGroot (1970). Because this source has proved error-free, I belive that these results can be used as is: they are "innocent until proven guilty." All of these results assume that the multivariate normal process is k-dimensional ($k \geq 1$) nonsingular, that is, the prescision matrices are assumed to be symetric positive definite.

5.3.1 Unknown Mean Vector, Known Precision Matrix

Suppose that data $\mathbf{x}_1, ..., \mathbf{x}_n$ are independent and identically disrtibuted from a k-dimensional ($k \geq 1$) multivariate normal distribution where the mean vector, μ, is unknown and the precision matrix, ρ, is known. The likelihood function, $L(\mu \mid \mathbf{x}_1, ..., \mathbf{x}_n)$, is proportional to μ is

$$L(\mu \mid \mathbf{x}_1, ..., \mathbf{x}_n) \propto \exp\left(-\frac{1}{2}(\mu - \bar{\mathbf{x}})'(n\rho)(\mu - \bar{\mathbf{x}})\right). \tag{155}$$

The sufficient statisitics are n, the number of data vectors, and $\bar{\mathbf{x}}$, the mean vector of the data. As a function of μ, Equation 4.6 is proportional to a k-dimensional multivariate normal distribution. Consequently, the conjugate prior, $\pi(\cdot)$, is a multivariate normal distribution with hyperparmeters $\mathbf{m}$, the mean vector, and $\mathbf{p}$, the symmetric positive definite precision matrix

$$\pi(\mu \mid \mathbf{m}, \mathbf{p}) = (2\pi)^{-k/2} |\mathbf{p}|^{1/2} \exp\left(-\frac{1}{2}(\mu - \mathbf{m})' \mathbf{p} (\mu - \mathbf{m})\right). \tag{156}$$

The posterior distribution, $\pi(\mu \mid \mathbf{m}^*, \mathbf{p}^*)$, is a k-dimensional multivariate normal distribution with hyperparameters

$$\mathbf{m}^* = (\mathbf{p} + n\rho)^{-1}(\mathbf{pm} + n\rho\bar{\mathbf{x}}) \quad \mathbf{p}^* = \mathbf{p} + n\rho. \tag{157}$$

5.3.2 Unknown Precision Matrix, Known Mean Vector

Suppose that data $\mathbf{x}_1, ..., \mathbf{x}_n$ are independent and identically disrtibuted from a k-dimensional ($k \geq 1$) multivariate normal distribution where the mean vector, μ, is known and the precision matrix, ρ, is unknown. The likelihood function, $L(\rho \mid \mathbf{x}_1, ..., \mathbf{x}_n)$, is proportional to ρ is

$$L(\rho \mid \mathbf{x}_1, ..., \mathbf{x}_n) \propto |\rho|^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)' \rho (\mathbf{x}_i - \mu)\right). \tag{158}$$

Since the exponent in Equation 4.above is a real number, a 1 x 1 matrix, we make use of the following relationship:

$$\sum_{i=1}^{n}(\mathbf{x}_i - \mu)' \rho (\mathbf{x}_i - \mu) = tr\left[\left(\sum_{i=1}^{n}(\mathbf{x}_i - \mu)(\mathbf{x}_i - \mu)'\right)\rho\right].$$

Making use of this relationship, the likelihood function can be rewritten as

$$L\left(\rho \mid \mathbf{x}_1, ..., \mathbf{x}_n\right) \propto |\rho|^{n/2} \exp\left(-\frac{1}{2} tr\left[\left(\sum_{i=1}^{n} \left(\mathbf{x}_i - \mu\right)\left(\mathbf{x}_i - \mu\right)'\right)\rho\right]\right). \tag{159}$$

The sufficient statistics are n, the number of data vectors, and $\sum_{i=1}^{n} \left(\mathbf{x}_i - \mu\right)\left(\mathbf{x}_i - \mu\right)'$. As a function of ρ, Equation 4.above is proportional to a Wishart distribution. Consequently, the conjugate prior, $\pi(\cdot)$, is a Wishart distribution with hyperparmeters $\alpha > k - 1$, the degrees of freedom, and τ, the symmetric positive definite precision matrix

$$\begin{aligned} \pi(\rho \quad | \quad \alpha, \tau) &= K\,|\tau|^{\alpha/2}\,|\rho|^{(\alpha-k-1)/2} \exp\left(-\frac{1}{2} tr\left(\tau\rho\right)\right) \\ where \quad K^{-1} \quad &= \quad 2^{\alpha k/2} \pi^{k(k-1)/4} \prod_{j=1}^{k} \Gamma\left(\frac{\alpha+1-j}{2}\right). \end{aligned} \tag{160}$$

The posterior distribution, $\pi(\rho \mid \alpha^*, \tau^*)$, is a Wishart distribution with hyperparameters

$$\tau^* = \tau + \sum_{i=1}^{n} \left(\mathbf{x}_i - \mu\right)\left(\mathbf{x}_i - \mu\right)' \quad \alpha^* = \alpha + n. \tag{161}$$

5.3.3 Unknown Mean Vector and Precision Matrix

Suppose that data $\mathbf{x}_1, ..., \mathbf{x}_n$ are independent and identically disrtibuted from a k-dimensional ($k \geq 1$) multivariate normal distribution where the mean vector, μ, and the precision matrix, ρ, are unknown. The likelihood function, $L\left(\mu, \rho \mid \mathbf{x}_1, ..., \mathbf{x}_n\right)$, proportional to the parameters is

$$L\left(\mu, \rho \mid \mathbf{x}_1, ..., \mathbf{x}_n\right) \propto |\rho|^{n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left(\mathbf{x}_i - \mu\right)' \rho \left(\mathbf{x}_i - \mu\right)\right). \tag{162}$$

The argument of the exponent above is a function of both unknown parameters. Using the following identities it is possible to untangle the dependence between these two pareameters and reexpress the likelihood function in more convenient form. First note that

$$\begin{aligned} &\sum_{i=1}^{n} \left(\mathbf{x}_i - \mu\right)' \rho \left(\mathbf{x}_i - \mu\right) \\ = \quad & n\left(\mu - \bar{\mathbf{x}}\right)' \rho \left(\mu - \bar{\mathbf{x}}\right) + \sum_{i=1}^{n} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)' \rho \left(\mathbf{x}_i - \bar{\mathbf{x}}\right). \end{aligned} \tag{163}$$

Second, we can further manipulate the second term in the above expression as

$$\sum_{i=1}^{n} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)' \rho \left(\mathbf{x}_i - \bar{\mathbf{x}}\right) = tr\left[\left(\sum_{i=1}^{n} \left(\mathbf{x}_i - \bar{\mathbf{x}}\right)\left(\mathbf{x}_i - \bar{\mathbf{x}}\right)'\right)\rho\right]. \tag{164}$$

Using these results, one can see that the likelihood function has one exponential component that is a function of both μ and ρ, while the other exponential component is a function of ρ alone:

$$
\begin{aligned}
& L\left(\mu,\rho \mid \mathbf{x}_1,...,\mathbf{x}_n\right) \\
\propto\ & |\rho|^{n/2}\exp\left(-\frac{n}{2}\left(\mu-\bar{\mathbf{x}}\right)'\rho\left(\mu-\bar{\mathbf{x}}\right)\right) \\
& \times\exp\left(-\frac{1}{2}tr\left[\left(\sum_{i=1}^{n}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)'\right)\rho\right]\right). \qquad (165)
\end{aligned}
$$

The sufficient statistics are n, the number of data vectors, $\bar{\mathbf{x}}$, the mean vector of the data, and $\sum_{i=1}^{n}\left(\mathbf{x}_i-\mu\right)\left(\mathbf{x}_i-\mu\right)'$, the sample covariance matrix of the data. Thus, the likelihood function is proportional to the product of a multivariate normal distribtution of μ given ρ and a Wishart distribution of ρ. Consequently, the conjugate prior, $\pi(\cdot,\cdot)$, is a multivariate normal-Wishart distribution. The Wishart hyperparmeters for the conjugate prior are $\alpha>k-1$, the degrees of freedom, and τ, the symmetric positive definite precision matrix and the hyperparameters that specify the Multivariate Normal component of the prior are $\mathbf{m}$, the mean vector, and $t\rho$, the symmetric positive definite precision matrix where $t>0$:

$$
\pi(\mu,\rho \mid \mathbf{m},t,\alpha,\tau)=\left(\frac{2\pi}{t}\right)^{-k/2}|\rho|^{1/2}\exp\left(-\frac{t}{2}\left(\mu-\mathbf{m}\right)'\rho\left(\mu-\mathbf{m}\right)\right)
$$

$$
\times K\left|\tau\right|^{\alpha/2}|\rho|^{(\alpha-k-1)/2}\exp\left(-\frac{1}{2}tr\left(\tau\rho\right)\right) \qquad (166)
$$

$$
where \quad K^{-1} = 2^{\alpha k/2}\pi^{k(k-1)/4}\prod_{j=1}^{k}\Gamma\left(\frac{\alpha+1-j}{2}\right). \qquad (167)
$$

The posterior distribution, $\pi(\mu,\rho \mid \mathbf{m}^*,t^*,\alpha^*,\tau^*)$, is a Multivariate Normal-Wishart distribution with hyperparameters

$$
\begin{aligned}
\mathbf{m}^* &= (t+n)^{-1}\left(\mathbf{tm}+n\bar{\mathbf{x}}\right) & t^* &= t+n \\
\tau^* &= \tau+\sum_{i=1}^{n}\left(\mathbf{x}_i-\mu\right)\left(\mathbf{x}_i-\mu\right)'+\tfrac{tn}{t+n}\left(\mathbf{m}-\bar{\mathbf{x}}\right)\left(\mathbf{m}-\bar{\mathbf{x}}\right)' & \alpha^* &= \alpha+n.
\end{aligned}
$$

DeGroot (pp. 179, 1970) shows that when the joint distribution of μ and ρ is a multivariate normal-wishart distribution the marginal distribution of the mean vector, μ, is a multivaritate t distribution. This distribution has $\alpha-k+1$ degrees of freedom, location vector $\mathbf{m}$, and precision matrix $t\left(\alpha-k+1\right)\tau^{-1}$.

$$
\begin{aligned}
\pi\left(\mu\right) &= C\left[1+\frac{1}{\alpha-k+1}\left(\mu-\mathbf{m}\right)'\left(t\left(\alpha-k+1\right)\tau^{-1}\right)\left(\mu-\mathbf{m}\right)\right]^{-(\alpha+1)/2}, \\
where \quad C &= \frac{\Gamma\left[\left(\alpha+1\right)/2\right]|T|^{1/2}}{\Gamma\left[\left(\alpha-k+1\right)/2\right]\left[\left(\alpha-k+1\right)\pi\right]^{k/2}}.
\end{aligned}
$$

5.3.4 Unknown Mean Vector and Precision Matrix known only up to its Scale Factor

Suppose that data $\mathbf{x}_1, ..., \mathbf{x}_n$ are independent and identically disrtibuted from a k-dimensional ($k \geq 1$) multivariate normal distribution where the mean vector, μ, is known and the precision matrix, $w\rho$, is only known up to scale factor, w. The likelihood function, $L(\mu,w \mid \mathbf{x}_1, ..., \mathbf{x}_n)$, proportional to the parameters is

$$L(\mu,w \mid \mathbf{x}_1, ..., \mathbf{x}_n) \propto |w\rho|^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \mu)'(w\rho)(\mathbf{x}_i - \mu)\right). \tag{168}$$

The likelihood function can be factored using the identities (4.15) and (4.16) to yield one factor proportional to the multivariate normal distribution of μ given w and one factor proportional to a gamma distribution of w:

$$\begin{aligned} & L(\mu,w \mid \mathbf{x}_1, ..., \mathbf{x}_n) \\ \propto\ & |w\rho|^{n/2} \exp\left(-\frac{n}{2}(\mu - \bar{\mathbf{x}})'(w\rho)(\mu - \bar{\mathbf{x}})\right) \\ & \times \exp\left(-\frac{w}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'\rho(\mathbf{x}_i - \bar{\mathbf{x}})\right). \end{aligned} \tag{169}$$

The sufficient statistics are n, the number of data vectors, $\bar{\mathbf{x}}$, the mean vector of the data, and $\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'\rho(\mathbf{x}_i - \bar{\mathbf{x}})$. The conjugate prior, $\pi(\cdot,\cdot)$, is a multivariate normal-gamma distribution. The h yperparmeters that specify the gamma of the conjugate prior are the shape parameter, $\alpha > 0$, and the scale parameter, $\beta > 0$. The hyperparameters that specify the multivariate normal component of the prior are $\mathbf{m}$, the mean vector, and $w\rho$, the scaled precision matrix. The conjugate prior is

$$\begin{aligned} \pi(\mu,w \mid \mathbf{m}, \mathbf{p},\alpha,\beta) = (2\pi)^{-k/2}|w\mathbf{p}|^{1/2} \exp\left(-\frac{1}{2}(\mu - \mathbf{m})'(w\mathbf{p})(\mu - \mathbf{m})\right) \\ \times\frac{\beta^\alpha}{\Gamma(\alpha)}w^{\alpha-1}\exp(-\beta w) \end{aligned} \tag{170}$$

The posterior distribution, $\pi(\mu,w \mid \mathbf{m}^*, \mathbf{p}^*,\alpha^*,\beta^*)$, is a Multivariate Normal-gamma distribution with hyperparameters

$$\begin{aligned} \mathbf{m}^* &= (\mathbf{p}+n\rho)^{-1}(\mathbf{pm}+n\rho\bar{\mathbf{x}}) & \mathbf{p}^* &= w(\mathbf{p}+n\rho) \\ \beta^* &= \beta+\tfrac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})'\rho(\mathbf{x}_i - \bar{\mathbf{x}}) + \tfrac{1}{2}(\mathbf{m}^* - \mathbf{m})'\mathbf{p}(\bar{\mathbf{x}} - \mathbf{m}) & \alpha^* &= \alpha+\tfrac{nk}{2}. \end{aligned}$$

The marginal distribution of the mean vector, μ, is a multivariate t distribution with 2α degrees of freedom, location vector μ, and precision matrix $(\alpha/\beta)\mathbf{p}$.

5.4 Continuous Data: Normal Regression Process

The results in this section come from Raiffa and Schlaifer (1961). This source has proved error-free and can be trusted as such.

We define the Normal regression process as the generator of data $y_1, ..., y_n$ such that for any $i = 1..n$ the density of y_i is

$$f(y_i) = \sqrt{\frac{\rho}{2\pi}} \exp\left[\frac{-\rho}{2}\left(y_i - \sum_{j=1}^{r} x_{ij}\beta_i\right)\right] \tag{171}$$

where the x_{ij} 's are known. In an attempt to simplify the notation we define the vectors β, $\mathbf{y}$, and $\mathbf{X}$:

$$\beta = (\beta_1 ... \beta_r)' \quad \mathbf{y} = (y_1 ... y_n)'$$

$$\mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1r} \\ \vdots & \ddots & & & \vdots \\ x_{i1} & & x_{ij} & & x_{ir} \\ \vdots & & & \ddots & \vdots \\ x_{n1} & \cdots & x_{nj} & \cdots & x_{nr} \end{bmatrix}.$$

5.4.1 Known Common Precision, Unknown Regression Coefficients

Suppose that data $y_1, ..., y_n$ are generated by a normal regression process where the common precision, ρ, is known and the vector of regression coefficients, β, is unknown. The likelihood function, $L(\beta \mid \mathbf{y})$, proportional to β is

$$L(\beta \mid \mathbf{y}) \propto \exp\left(-\frac{\rho}{2}(\mathbf{b} - \beta)' \mathbf{n} (\mathbf{b} - \beta)\right), \tag{172}$$

where $\mathbf{n} = \mathbf{X}'\mathbf{X}$ and $\mathbf{b}$ satisfies $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$.

The sufficient statisitics are $\mathbf{n}$, the normal equations, and $\mathbf{b}$, the least square coefficients. As a function of β, the likelihood is proportional to a r-dimensional multivariate normal distribution. Consequently, the conjugate prior, $\pi(\cdot)$, is a multivariate normal distribution with hyperparmeters $\mathbf{m}$, the mean vector, and $\mathbf{p}$, the symmetric positive definite precision matrix

$$\pi(\beta \mid \mathbf{m}, \rho\mathbf{p}) = \left(\frac{\rho}{2\pi}\right)^{r/2} |\mathbf{p}|^{1/2} \exp\left(-\frac{\rho}{2}(\mathbf{m} - \beta)' \mathbf{p} (\mathbf{m} - \beta)\right). \tag{173}$$

The posterior distribution, $\pi(\beta \mid \mathbf{m}^*, \rho\mathbf{p}^*)$, is a r-dimensional multivariate normal distribution with hyperparameters

$$\mathbf{m}^* = (\mathbf{p} + \mathbf{n})^{-1} (\mathbf{pm} + \mathbf{nb}) \quad \mathbf{p}^* = \mathbf{p} + \mathbf{n}. \tag{174}$$

5.4.2 Unknown Common Precision and Regression Coefficients

Suppose that data $y_1, ..., y_n$ are generated by a normal regression process where the common precision, ρ, and the vector of regression coefficients, β, are unknown. The likelihood function, $L(\beta, \rho \mid \mathbf{y})$, proportional to β and ρ is

$$L(\beta, \rho \mid \mathbf{y}) \propto \rho^{\nu/2} \exp\left(\frac{-1}{2}\rho\nu v\right) \exp\left(-\frac{\rho}{2}(\mathbf{b} - \beta)' \mathbf{n} (\mathbf{b} - \beta)\right),$$

where $\mathbf{n} = \mathbf{X}'\mathbf{X}$, $\mathbf{b}$ satisfies $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{y}$, n, the number of data points, $\nu = n - rank(\mathbf{n})$, and $v = \frac{1}{\nu}(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$.

The sufficient statisitics are $\mathbf{n}$, the normal equations, and $\mathbf{b}$, the least square coefficients, ν, and ν . As a function of the unknown parameters the likelihood is proportional to a multivariate normal-gamma distribution. The multivariate normal-gamma conjugate prior is specified by hyperparmeters $\alpha, \beta > 0$, $\mathbf{m}$, the mean vector, and , $\mathbf{p}$ the symmetric postitive definitie precision matrix:

$$\pi(\beta,\rho \mid \mathbf{m}, \mathbf{p},\alpha, \beta) = \left(\frac{\rho}{2\pi}\right)^{r/2} |\mathbf{p}|^{1/2} \exp\left(-\frac{\rho}{2}(\beta - \mathbf{m})' \mathbf{p}(\beta - \mathbf{m})\right)$$
$$\times \frac{\left(\frac{\alpha\beta}{2}\right)^{\alpha/2}}{\Gamma\left(\frac{\alpha}{2}\right)} \rho^{\alpha/2-1} \exp\left(-\frac{\alpha\beta}{2}\rho\right). \tag{175}$$

The posterior distribution, $\pi(\beta,\rho \mid \mathbf{m}^*, \mathbf{p}^*,\alpha^*, \beta^*)$, is a multivariate normal-gamma distribution with hyperparameters

$$\begin{array}{cc} \mathbf{m}^* = (\mathbf{p} + \mathbf{n})^{-1}(\mathbf{pm} + \mathbf{np}) & \mathbf{p}^* = \mathbf{p} + \mathbf{n} \\ \beta^* = \frac{1}{\alpha+n}\left[(\alpha\beta + \mathbf{m}'\mathbf{pm}) + (\nu v + \mathbf{b}'\mathbf{nb}) + \mathbf{m}^{*\prime}\mathbf{p}^*\mathbf{m}^*\right] & \alpha^* = \alpha+n. \end{array}$$

The marginal distribution of the mean vector, β, is a multivariate t distribution with ν degrees of freedom, location vector $\mathbf{b}$, and precision matrix $\frac{1}{v}\mathbf{n}$.

6 Jeffreys's Priors

Although Jeffreys's (Jeffreys's 1961) priors are rarely conjugate priors, they often give rise to analytically tractable posterior distributions. Moreover, the posterior distributions arising from Jeffrey's priors are often members of the conjugate family for the likelihood function.. The priors in this table are based on Hartigan (1964).

Table 1: Jeffreys's Priors for Several Data Generating Processes

Data Generating Process		Unknown Parameter(s)
Binomial	$f(x \mid p) = nxp^x(1-p)^{n-x}$	p
Negative Binomial	$f(x \mid p) = r-1x-1p^r(1-p)^{x-r}$	p
Poisson	$f(x \mid \mu) = \frac{\mu^x \exp(-\mu)}{x!}$	μ
Normal	$f(x \mid \mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2}(x-\mu)^2\right)$	μ, mean
Normal	$f(x \mid \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(x-\mu)^2\right)$	σ, standard deviation
Normal	$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2\sigma^2}(x-\mu)^2\right)$	μ, σ mean and sd
Gamma	$f(x \mid \alpha) = \frac{x^{\alpha-1} \exp\frac{-x}{\beta}}{\Gamma(\alpha)\beta^\alpha}$	α, shape parameter
Multinomial	$f(\mathbf{x} \mid \theta_1 \cdots \theta_r) \propto \frac{n!}{x_1! \cdots x_r!} \theta_1^{x_1} \cdots \theta_r^{x_r}$	$\theta_1 \cdots \theta_r$
n-dimensional Normal	$f(\mathbf{x} \mid \mu) \propto (2\pi)^{-n/2} \exp\left[\frac{-1}{2}(\mathbf{x}-\mathbf{K}\mu)'(\mathbf{x}-\mathbf{K}\mu)\right]$	μ, mean vector in R^k
n-dimensional Normal	$f(\mathbf{x} \mid \mathbf{\Sigma}) \propto \lvert\mathbf{\Sigma}\rvert^{1/2} (2\pi)^{-n/2} \exp\left[\frac{-1}{2}\mathbf{x}'\mathbf{\Sigma}\mathbf{x}\right]$	$\mathbf{\Sigma}$, Covariance matrix

7 References

References

[1] Athreya, K. B. "Another Conjugate Family for the Normal Distribution" *Statistics & Probability Letters*, 1986, **4**, 61-64

[2] Arnold, Barry C. and Press, S. James. "Bayesian Estimation and Prediction for Pareto Data", *Journal of the American Statistical Associatio*n 1989, **84**, 1079-1084

[3] Banjerjee, Asit K. and Bhattacharyya, G. K. "Bayesian Results for the Inverse Gaussian Distribution with an Application" *Technometrics*, 1979, **21**(2), 247-251.

[4] Box,G.E.P. and Tiao,G.C. *Bayesian Inference in Statistical Analysis.* 1973, Addison-Wesley, Reading, Mass.

[5] Dalal, S. and Hall, W. J. "Approximating Priors by Mixtures of Natural Conjugate Priors" *J. R. Statist. Soc.*, **B,** 1983, **45**, 278-286.

[6] Darmois, G. "Sur les lois de probabilit estimation exhaustive." *C. R. Acad. Sci. Paris*, 1935, **260**, 1265-1266

[7] DeGroot, Morris H. *Optimal Statistical Decisions*, 1970, McGraw-Hill Book Company, New York

[8] Diaconis, Perci and Ylvsaker, Donald "Conjugate Priors for Exponential Families" *The Annals of Statistics* 1979, **7**(2),. 269-281

[9] Diaconis, Perci and Ylvsaker, Donald "Quantifying prior Opinion." In: *Bayesian Statistics 2*, J.M. Bernardo et al. (ed.), 1985, 133-156. North-Holland: Amsterdam.

[10] Dickey, James M. "The Weighted Likelihood Ratio, Linear Hypotheses on Normal Location Parameters" *The Annals of Mathematical Statistics*,. 1971, **42**(1), 204-223

[11] Dyer, Danny and Pierce, Rebecca L. "On the Choice of the Prior Distribution in Hypergeometric Sampling". *Communications in Statistics-Theory and Methods*, 1993, **22**(8), 2125-2146

[12] Guttorp, Peter and Lockhart, Richard A. "Finding the Location of a Signal:A Bayesian Approach", *Journal of the American Statistical Association* 1988, **83**, 322-330

[13] Gradshteyn, I.S. and Ryzhik, I.M.; Jeffrey, Alan editor. *Table of Integrals, Series, and Products, 5th edition.* 1994 Academic Press, San Diego

[14] Hartigan, J. (1964). Invariant prior distributions. *Annals of Mathematical Statistics*, **35**, 836-845

[15] Jeffreys, H. (1961). *Theory of Probability*, 3rd ed. Clarendon Press, Oxford (1st. ed , 1939)

[16] Koopman, B. O. "On distributions admitting a sufficient statistic", 1936, *Trans. Am. Math. Soc.*, **39**, 399-409

[17] LaValle, Irving H. *An Introduction to Probability, Decision, and Inference*, 1970 Holt, Rinehart and Winston, Inc. New York

[18] Mardia, K.V. and El-Atoum, S.A.M. "Bayesian inference for the von Mises-Fisher distribution" *Biometrika*, 1976, **63**, 203-206

[19] Miller, Robert B. "Bayesian Analysis of the Two-Parameter Gamma Distribution" *Technometrics*, 1980, **22**(1), 65-69

[20] O'Hagan, Anthony *Kendall's Advanced Theory of Statistics, Vol. 2B Bayesian Inference* 1994 Halsted Press, New York.

[21] Pitman, E. J. G. "Sufficient statistics and intrinsic accuracy". *Proc. Cambridge Phil. Soc.*, 1936, **32**, 567-579

[22] Raiffa, Howard and Schlaifer, Robert *Applied Statistical Decision Theory, Division of Research*, 1961 Graduate School of Business Administration, Harvard University, Boston.

[23] Smith,A.F.M. and Gelfand,A.E. "Bayesian Statistics without Tears: A Sampling-Resampling Perspective" 1992, *The American Statistician*, **46**, 84-88