3. Conjugate families of distributions

Objective

One problem in the implementation of Bayesian approaches is analytical tractability. For a likelihood function $l(\boldsymbol{\theta}|\mathbf{x})$ and prior distribution $p(\boldsymbol{\theta})$, in order to calculate the posterior distribution, it is necessary to evaluate

$$f(\mathbf{x}) = \int l(\boldsymbol{\theta}|\mathbf{x})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

and to make predictions, we must evaluate

$$f(y|\mathbf{x}) = \int f(y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{x})\, d\boldsymbol{\theta}.$$

General approaches for evaluating such integrals numerically are studied in chapter 6, but in this chapter, we will study the conditions under which such integrals may be evaluated analytically.

Recommended reading

- Wikipedia entry on conjugate priors.

  http://en.wikipedia.org/wiki/Conjugate_prior

- Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory*, Section 5.2.

Coin tossing problems and beta priors

It is possible to generalize what we have seen in Chapter 2 to other coin tossing situations.

Suppose that we have defined a beta prior distribution $\theta \sim \mathcal{B}(\alpha, \beta)$ for $\theta = P(\text{head})$.

Then if we observe a sample of coin toss data, whether the sampling mechanism is binomial, negative-binomial or geometric, the likelihood function always takes the form

$$l(\theta|\mathbf{x}) = c\theta^h(1-\theta)^t$$

where c is some constant that depends on the sampling distribution and h and t are the observed numbers of heads and tails respectively.

Applying Bayes theorem, the posterior distribution is

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto c\theta^h(1-\theta)^t \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{\alpha+h-1}(1-\theta)^{\beta+t-1} \end{aligned}$$

which implies that $\theta|\mathbf{x} \sim \mathcal{B}(\alpha+h, \beta+t)$.

Thus, using a beta prior, guarantees that the posterior distribution is also beta. In this case, we say that the class of beta prior distributions is $conjugate$ to the class of binomial (or geometric or negative binomial) likelihood functions.

Interpretation of the beta parameters

The posterior distribution is $\mathcal{B}(\alpha + h, \beta + t)$ so that α (β) in the prior plays the role of the number of heads (tails) observed in the experiment. The information represented by the prior distribution can be viewed as equivalent to the information contained in an experiment where we observe α heads and β tails.

Furthermore, the posterior mean is given by $(\alpha + h)/(\alpha + \beta + h + t)$ which can be expressed in mixture form as

$$\begin{aligned} E[\theta|\mathbf{x}] &= w\frac{\alpha}{\alpha+\beta} + (1-w)\frac{h}{h+t} \\ &= wE[\theta] + (1-w)\hat{\theta} \end{aligned}$$

where $w = \frac{\alpha+\beta}{\alpha+\beta+h+t}$ is the relative weight of the number of equivalent coin tosses in the prior.

Limiting results

Consider also what happens as we let $\alpha, \beta \to 0$. We can interpret this as representing a prior which contains information equivalent to zero coin tosses.

In this case, the posterior distribution tends to $\mathcal{B}(h, t)$ and, for example, the posterior mean $E[\theta|\mathbf{x}] \to \hat{\theta}$, the classical MLE.

The limiting form of the prior distribution in this case is Haldane's (1931) prior,

$$p(\theta) \propto \frac{1}{\theta(1-\theta)},$$

which is an *improper* density. We analyze the use of improper densities further in chapter 5.

Prediction and posterior distribution

The form of the predictive and posterior distributions depends on the sampling distribution. We shall consider 4 cases:

- Bernoulli trials.
- Binomial data.
- Geometric data.
- Negative binomial data.

Bernoulli trials

Given that the beta distribution is conjugate in coin tossing experiments, given a (Bernoulli or binomial, etc.) sampling distribution, $f(x|\theta)$, we only need to calculate the prior predictive density $f(x) = \int f(x|\theta)p(\theta)\,d\theta$ for a beta prior. Assume first that we have a Bernoulli trial with $P(X = 1|\theta) = \theta$ and prior distribution $\theta \sim \mathcal{B}(\alpha, \beta)$. Then:

Theorem 3

If X is a Bernoulli trial with parameter θ and $\theta \sim \mathcal{B}(\alpha, \beta)$, then:

$$f(x) = \begin{cases} \frac{\alpha}{\alpha+\beta} & \text{if } x = 1 \\ \frac{\beta}{\alpha+\beta} & \text{if } x = 0 \end{cases}$$

$$E[X] = \frac{\alpha}{\alpha+\beta}, \qquad V[X] = \frac{\alpha\beta}{(\alpha+\beta)^2}.$$

Given a sample $\mathbf{x} = (x_1, \ldots, x_n)$ of Bernoulli data, the posterior distribution is $\theta|\mathbf{x} \sim \mathcal{B}(\alpha + \sum_{i=1}^n x_i, \beta + n - \sum_{i=1}^n x_i)$.

Proof

$$
\begin{aligned}
P(X=1) &= \int_0^1 P(X=1|\theta)p(\theta)\,d\theta \\
&= E[\theta] = \frac{\alpha}{\alpha+\beta} \\
P(X=0) &= 1 - P(X=1) = \frac{\beta}{\alpha+\beta} \\
E[X] &= 0 \times P(X=0) + 1 \times P(X=1) = \frac{\alpha}{\alpha+\beta} \\
V[X] &= E\left[X^2\right] - E[X]^2 \\
&= \frac{\alpha}{\alpha+\beta} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 \\
&= \frac{\alpha\beta}{(\alpha+\beta)^2}.
\end{aligned}
$$

The posterior distribution formula was demonstrated previously. ∎

Theorem 4

Suppose that $X|\theta \sim \mathcal{BI}(m, \theta)$ with $\theta \sim \mathcal{B}(\alpha, \beta)$. Then, the predictive density of X is the beta binomial density

$$\begin{aligned} f(x) &= \binom{m}{x} \frac{B(\alpha + x, \beta + m - x)}{B(\alpha, \beta)} \quad \text{for } x = 0, 1, \ldots, m \\ E[X] &= m\frac{\alpha}{\alpha + \beta} \\ V[X] &= m(m + \alpha + \beta)\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \end{aligned}$$

Given a sample, $\mathbf{x} = (x_1, \ldots, x_n)$, of binomial data, the posterior distribution is $\theta|\mathbf{x} \sim \mathcal{B}(\alpha + \sum_{i=1}^n x_i, \beta + mn - \sum_{i=1}^n x_i)$.

Proof

$$
\begin{aligned}
f(x) &= \int_0^1 f(x|\theta)p(\theta)\,d\theta \\
&= \int_0^1 \binom{m}{x} \theta^x(1-\theta)^{m-x}\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta \quad \text{for } x = 0,\ldots,m \\
&= \binom{m}{x}\frac{1}{B(\alpha,\beta)}\int_0^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+m-x-1}\,d\theta \\
&= \binom{m}{x}\frac{B(\alpha+x,\beta+m-x)}{B(\alpha,\beta).}
\end{aligned}
$$

$$
\begin{aligned}
E[X] &= E[E[X|\theta]] \\
&= E[m\theta] \quad \text{remembering the mean of a binomial distribution} \\
&= m\frac{\alpha}{\alpha+\beta}.
\end{aligned}
$$

$$
\begin{aligned}
V[X] &= E[V[X|\theta]] + V[E[X|\theta]] \\
&= E[m\theta(1-\theta)] + V[m\theta] \\
&= m\frac{B(\alpha+1,\beta+1)}{B(\alpha,\beta)} + m^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= m\frac{\alpha\beta}{(\alpha+\beta)(\alpha+\beta+1)} + m^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \\
&= m(m+\alpha+\beta)\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.
\end{aligned}
$$

Theorem 5
Suppose that $\theta \sim \mathcal{B}(\alpha, \beta)$ and $X|\theta \sim \mathcal{GE}(\theta)$, that is

$$P(X = x|\theta) = (1-\theta)^x \theta \quad \text{for } x = 0, 1, 2, \ldots$$

Then, the predictive density of X is

$$\begin{aligned} f(x) &= \frac{B(\alpha+1, \beta+x)}{B(\alpha,\beta)} \quad \text{for } x = 0, 1, 2, \ldots \\ E[X] &= \frac{\beta}{\alpha - 1} \quad \text{for } \alpha > 1 \\ V[X] &= \frac{\alpha\beta(\alpha+\beta-1)}{(\alpha-1)^2(\alpha-2)} \quad \text{for } \alpha > 2. \end{aligned}$$

Given a sample, $\mathbf{x} = (x_1, \ldots, x_n)$, then $\theta|\mathbf{x} \sim \mathcal{B}(\alpha + n, \beta + \sum_{i=1}^n x_i)$.

Proof Exercise. ∎

Theorem 6

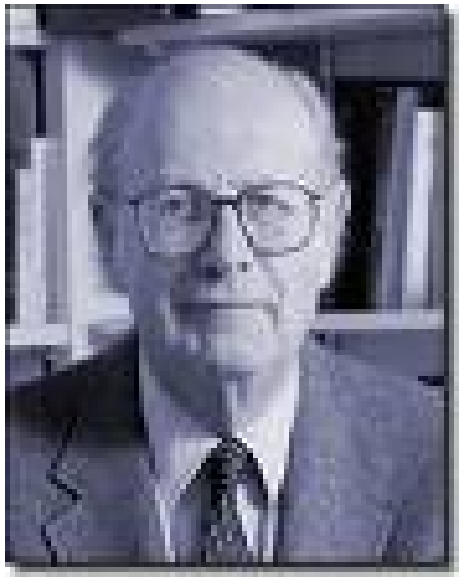
Suppose that $\theta \sim \mathcal{B}(\alpha, \beta)$ and $X|\theta \sim \mathcal{NB}(r, \theta)$, that is $P(X = x|\theta) = \begin{pmatrix} x+r-1 \\ x \end{pmatrix} \theta^r (1-\theta)^x$ for $x = 0, 1, 2, \ldots$ Then:

$$\begin{aligned} f(x) &= \begin{pmatrix} x+r-1 \\ x \end{pmatrix} \frac{B(\alpha+r, \beta+x)}{B(\alpha, \beta)} \quad \text{for } x = 0, 1, \ldots \\ E[X] &= r\frac{\beta}{\alpha-1} \quad \text{for } \alpha > 1 \\ V[X] &= \frac{r\beta(r+\alpha-1)(\alpha+\beta-1)}{(\alpha-1)^2(\alpha-2)} \quad \text{for } \alpha > 2. \end{aligned}$$

Given a sample $\mathbf{x} = (x_1, \ldots, x_n)$ of negative binomial data, the posterior distribution is $\theta|\mathbf{x} \sim \mathcal{B}(\alpha + rn, \beta + \sum_{i=1}^n x_i)$.

Proof Exercise. ∎

Conjugate priors



Raiffa

The concept and formal definition of conjugate prior distributions comes from Raiffa and Schlaifer (1961).

Definition 2

If $\mathcal{F}$ is a class of sampling distributions $f(x|\boldsymbol{\theta})$ and $\mathcal{P}$ is a class of prior distributions for $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$ we say that $\mathcal{P}$ is **conjugate to** $\mathcal{F}$ if $p(\boldsymbol{\theta}|x) \in \mathcal{P} \; \forall \; f(\cdot|\boldsymbol{\theta}) \in \mathcal{F} \; y \; p(\cdot) \in \mathcal{P}$.

The exponential-gamma system

Suppose that $X|\theta \sim \mathcal{E}(\theta)$, is exponentially distributed and that we use a gamma prior distribution, $\theta \sim \mathcal{G}(\alpha, \beta)$, that is

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \quad \text{for} \;\; \theta > 0.$$

Given sample data $\mathbf{x}$, the posterior distribution is

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto p(\theta) l(\theta|\mathbf{x}) \\ &\propto \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \prod_{i=1}^{n} \theta e^{-\theta x_i} \\ &\propto \theta^{\alpha+n-1} e^{-(\beta+n\bar{x})\theta} \end{aligned}$$

which is the nucleus of a gamma density $\theta|\mathbf{x} \sim \mathcal{G}(\alpha+n, \beta+n\bar{x})$ and thus the gamma prior is conjugate to the exponential sampling distribution.

Interpretation and limiting results

The information represented by the prior distribution can be interpreted as being equivalent to the information contained in a sample of size α and sample mean β/α.

Letting $\alpha, \beta \to 0$, then the posterior distribution approaches $\mathcal{G}(n, n\bar{x})$ and thus, for example, the posterior mean tends to $1/\bar{x}$ which is equal to the MLE in this experiment. However, the limiting prior distribution in this case is

$$f(\theta) \propto \frac{1}{\theta}$$

which is improper.

Prediction and posterior distribution

Theorem 7
Let $X|\theta \sim \mathcal{E}(\theta)$ and $\theta \sim \mathcal{G}(\alpha, \beta)$. Then the predictive density of X is

$$\begin{aligned} f(x) &= \frac{\alpha\beta^\alpha}{(\beta+x)^{\alpha+1}} \quad \text{for } x > 0 \\ E[X] &= \frac{\beta}{\alpha-1} \quad \text{for } \alpha > 1 \\ V[X] &= \frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)} \quad \text{for } \alpha > 2. \end{aligned}$$

Given a sample $\mathbf{x} = (x_1, \ldots, x_n)$ of exponential data, the posterior distribution is $\theta|\mathbf{x} \sim \mathcal{G}(\alpha + n, \beta + n\bar{x})$.

Proof Exercise. ∎

Exponential families

The exponential distribution is the simplest example of an exponential family distribution. Exponential family sampling distributions are highly related to the existence of conjugate prior distributions.

Definition 3

A probability density $f(x|\theta)$ where $\theta \in \mathbb{R}$ is said to belong to the one-parameter $exponential\ family$ if it has form

$$f(x|\theta) = C(\theta)h(x)\exp(\phi(\theta)s(x))$$

for given functions $C(\cdot), h(\cdot), \phi(\cdot), s(\cdot)$. If the support of X is independent of θ then the family is said to be $regular$ and otherwise it is $irregular$.

Examples of exponential family distributions

Example 12

The binomial distribution,

$$\begin{aligned} f(x|\theta) &= \binom{m}{x} \theta^x (1-\theta)^{m-x} \quad \text{for } x = 0, 1, \ldots, m \\ &= (1-\theta)^m \binom{m}{x} \exp\left(x \log \frac{\theta}{1-\theta}\right) \end{aligned}$$

is a regular exponential family distribution.

Example 13

The Poisson distribution,

$$f(x|\theta) = \frac{\theta^x e^{-\theta}}{x!} = e^{-\theta} \frac{1}{x!} \exp(x \log \theta) \quad \text{for } x = 0, 1, 2, \ldots$$

is a regular exponential family distribution.

Irregular and non exponential family distributions

Example 14

The uniform distribution,

$$f(x|\theta) = \frac{1}{\theta} \quad \text{for } 0 < x < \theta$$

is an irregular exponential family distribution.

Example 15

The Cauchy density,

$$f(x|\theta) = \frac{1}{\pi\left(1 + (x-\theta)^2\right)} \quad \text{for} \quad -\infty < x < \infty$$

is not an exponential family distribution.

The Student's t and Fisher's F distributions and the logistic distribution are other examples of non exponential family distributions.

Sufficient statistics and exponential family distributions

Theorem 8

If $X|\theta$ is a one-parameter, regular exponential family distribution, then given a sample $\mathbf{x} = (x_1, \ldots, x_n)$, a sufficient statistic for θ is $t(\mathbf{x}) = \sum_{i=1}^{n} s(x_i)$.

Proof The likelihood function is

$$\begin{aligned} l(\theta|\mathbf{x}) &= \prod_{i=1}^{n} C(\theta)h(x_i)\exp\left(\phi(\theta)s(x_i)\right) \\ &= \left(\prod_{i=1}^{n} h(x_i)\right) C(\theta)^n \exp\left(\phi(\theta)t(\mathbf{x})\right) \\ &= h(\mathbf{x})g(t,\theta) \end{aligned}$$

where $h(\mathbf{x}) = \left(\prod_{i=1}^{n} h(x_i)\right)$ and $g(t,\theta) = C(\theta)^n \exp\left(\phi(\theta)t(\mathbf{x})\right)$ and therefore, t is a sufficient statistic as in Theorem 1. ∎

A conjugate prior to an exponential family distribution

If $f(x|\theta)$ is an exponential family, with density as in Definition 3, then a conjugate prior distribution for θ exists.

Theorem 9

The prior distribution $p(\theta) \propto C(\theta)^a \exp\left(\phi(\boldsymbol{\theta})b\right)$ is conjugate to the exponential family distribution likelihood.

Proof From Theorem 8, the likelihood function for a sample of size n is

$$l(\theta|\mathbf{x}) \propto C(\theta)^n \exp\left(\phi(\theta)t(\mathbf{x})\right)$$

and given $p(\boldsymbol{\theta})$ as above, we have a posterior of the same form

$$p(\theta|\mathbf{x}) \propto C(\theta)^{a^\star} \exp\left(\phi(\theta)b^\star\right)$$

where $a^\star = a + n$ and $b^\star = b + t(\mathbf{x})$ for $j = 1, \ldots, n$. ∎

The Poisson-gamma system

Suppose that $X|\theta \sim \mathcal{P}(\theta)$. Then, from Example 13, we have the exponential family form

$$f(x|\theta) = e^{-\theta}\frac{1}{x!}\exp(x\log\theta)$$

and from Theorem 9, a conjugate prior density for θ has form

$$p(\theta) \propto \left(e^{-\theta}\right)^{a}\exp(b\log\theta)$$

for some a, b. Thus, $p(\theta) \propto \theta^b e^{-a\theta}$ which is the nucleus of a gamma density, $\theta \sim \mathcal{G}(\alpha, \beta)$, where $\alpha = b + 1$ and $\beta = a$.

Thus, the gamma prior distribution is conjugate to the Poisson sampling distribution.

Derivation of the posterior distribution

Now assume the prior distribution $\theta \sim \mathcal{G}(\alpha, \beta)$ and suppose that we observe a sample of n Poisson data. Then, we can derive the posterior distribution via Bayes theorem as earlier.

$$\begin{aligned} p(\theta|\mathbf{x}) &\propto \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}\prod_{i=1}^{n}\frac{\theta^{x_i}e^{-\theta}}{x_i!} \\ &\propto \theta^{\alpha+n\bar{x}-1}e^{-(\beta+n)\theta} \\ \theta|\mathbf{x} &\sim \mathcal{G}(\alpha+n\bar{x}, \beta+n) \end{aligned}$$

Alternatively, we can use Theorem 9. The sufficient statistic is $t(\mathbf{x}) = \sum_{i=1}^{n} x_i$ and thus, from Theorem 9, we know immediately that

$$\theta|\mathbf{x} \sim \mathcal{G}(\underbrace{\alpha}_{b+1}+t(\mathbf{x}), \underbrace{\beta}_{a}+n) \sim \mathcal{G}(\alpha+n\bar{x}, \beta+n).$$

Interpretation and limiting results

We might interpret the information in the prior as being equivalent to a the information contained in a sample of size β with sample mean α/β.

Letting $\alpha, \beta \to 0$, the posterior mean converges to the MLE although the corresponding limiting prior, $p(\theta) \propto \frac{1}{\theta}$, is improper.

Prediction and posterior distribution

Theorem 10
Let $X|\theta \sim \mathcal{P}(\theta)$ with $\theta \sim \mathcal{G}(\alpha, \beta)$. Then:

$$\begin{aligned} X &\sim \mathcal{NB}\left(\alpha, \frac{\beta}{\beta+1}\right) \quad \text{that is,} \\ f(x) &= \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{x} \quad \text{for } x = 0, 1, 2, \ldots \\ E[X] &= \frac{\alpha}{\beta} \\ V[X] &= \frac{\alpha(\beta+1)}{\beta^2}. \end{aligned}$$

Given a sample $\mathbf{x} = (x_1, \ldots, x_n)$ of Poisson data, the posterior distribution is $\theta|\mathbf{x} \sim \mathcal{G}(\alpha + n\bar{x}, \beta + n)$.

Proof Exercise. ∎

The uniform-Pareto system

As noted in Example 14, the uniform distribution is not a regular exponential family distribution. However, we can still define a conjugate prior distribution.

Let $X \sim \mathcal{U}(0, \theta)$. Then, given a sample of size n, the likelihood function is $l(\theta|\mathbf{x}) = \frac{1}{\theta^n}, \quad \text{for } \theta > x_{\max}$, where $x_{\max}$ is the sample maximum.

Consider a $Pareto$ prior distribution, $\theta \sim \mathcal{PA}(\alpha, \beta)$, with density

$$p(\theta) = \frac{\alpha\beta^\alpha}{\theta^{\alpha+1}} \quad \text{for } \theta > \beta > 0$$

and mean $E[\theta] = \alpha\beta/(\alpha - 1)$.

It is clear that this prior distribution is conjugate and that

$$\theta|\mathbf{x} \sim \mathcal{PA}\left(\alpha^\star, \beta^\star\right)$$

where $\alpha^\star = \alpha + n$ and $\beta^\star = \max\{\beta, x_{\max}\}$.

Properties and limiting results

The posterior mean in this case is

$$E[\theta|\mathbf{x}] = \frac{(\alpha+n)\max\{\beta, x_{\max}\}}{\alpha+n-1}$$

which cannot be represented in a general way as an average of the prior mean and the MLE ($x_{\max}$).

We can interpret the information in the prior as being equivalent to a sample of size α where the maximum value is equal to β. When we let $\alpha, \beta \to 0$, then the posterior distribution approaches $\mathcal{PA}(n, x_{max})$ and in this case, the posterior mean is

$$E[\theta|\mathbf{x}] = \frac{nx_{\max}}{n-1} > x_{\max} = \hat{\theta}.$$

This also corresponds to an improper limiting prior $p(\theta) \propto 1/\theta$.

In this experiment, no prior distribution (except for a point mass at $x_{\max}$) will lead to a posterior distribution whose mean coincides with the MLE.

Prediction and posterior distribution

Theorem 11
If $X|\theta \sim \mathcal{U}(0,\theta)$ and $\theta \sim \mathcal{PA}(\alpha,\beta)$, then:

$$\begin{aligned} f(x) &= \begin{cases} \frac{\alpha}{(\alpha+1)\beta} & \text{if } 0 < x < \beta \\ \frac{\alpha\beta^{\alpha}}{(\alpha+1)x^{\alpha+1}} & \text{if } x \geq \beta \end{cases} \\ E[X] &= \frac{\alpha\beta}{2(\alpha-1)} \quad \text{for } \alpha > 1 \\ V[X] &= \frac{\alpha(\alpha+2)\beta^2}{12(\alpha-1)(\alpha-2)} \quad \text{for } \alpha > 2. \end{aligned}$$

Given a sample of size n, then the posterior distribution of θ is $\theta|\mathbf{x} \sim \mathcal{PA}\left(\alpha+n, \max\{\beta, x_{\max}\}\right)$.

Proof Exercise. ∎

Higher dimensional exponential family distributions

Definition 4

A probability density $f(\mathbf{x}|\boldsymbol{\theta})$ where $\theta \in \mathbb{R}^k$ is said to belong to the k-parameter *exponential family* if it has form

$$f(\mathbf{x}|\boldsymbol{\theta}) = C(\boldsymbol{\theta})h(\mathbf{x}) \exp\left(\sum_{j=1}^{k} \phi_j(\boldsymbol{\theta})s_j(\mathbf{x})\right)$$

for given functions $C(\cdot), h(\cdot), \phi(\cdot), s(\cdot)$. If the support of $\mathbf{X}$ is independent of θ then the family is said to be *regular* and otherwise it is *irregular*.

Example 16

The multinomial density is given by $f(\mathbf{x}|\boldsymbol{\theta}) = \frac{m!}{\prod_{j=1}^{k} x_j!} \prod_{j=1}^{k} \theta_j^{x_j}$ where $x_j \in \{0, 1, \ldots, m\}$ for $j = 1, \ldots, k$, $\sum_{j=1}^{k} x_j = m$ and $\sum_{j=1}^{k} \theta_j = 1$.

We can write

$$\begin{aligned} f(\mathbf{x}|\boldsymbol{\theta}) &= \frac{m!}{\prod_{j=1}^{k} x_j!} \exp\left(\sum_{j=1}^{k} x_j \log(\theta_j)\right) \\ &= \frac{m!}{\prod_{j=1}^{k} x_j!} \exp\left((m - \sum_{j=1}^{k-1} x_j) \log \theta_k + \sum_{j=1}^{k-1} x_j \log(\theta_j)\right) \\ &= \frac{m!}{\prod_{j=1}^{k} x_j!} \theta_k^m \exp\left(\sum_{j=1}^{k-1} x_j \log(\theta_j/\theta_k)\right) \end{aligned}$$

and thus, the multinomial distribution is a regular, $k-1$ dimensional exponential family distribution.

It is clear that we can generalize Theorem 8 to the k dimensional case.

Theorem 12
If $\mathbf{X}|\boldsymbol{\theta}$ is a k-parameter, regular exponential family distribution, then given a sample $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, a sufficient statistic for $\boldsymbol{\theta}$ is $\mathbf{t}(\mathbf{x}) = (\sum_{i=1}^n s_1(\mathbf{x}_i), \ldots, \sum_{i=1}^n s_k(\mathbf{x}_i))$.

Proof Using the same arguments as in Theorem 8, the result is easy to derive.

A conjugate prior to an exponential family sampling distribution

We can also generalize Theorem 9. If $f(x|\boldsymbol{\theta})$ is an exponential family, with density as in Definition 4, then a conjugate prior distribution for $\boldsymbol{\theta}$ exists.

Theorem 13
The prior distribution $p(\boldsymbol{\theta}) \propto C(\boldsymbol{\theta})^a \exp\left(\sum_{j=1}^{k} \phi_j(\boldsymbol{\theta}) b_j\right)$ is conjugate to the k dimensional exponential family distribution likelihood.

Proof Exercise. ∎

The multinomial-Dirichlet system

Suppose that $\mathbf{X}|\boldsymbol{\theta} \sim \mathcal{MN}(m, \boldsymbol{\theta})$ is a k- dimensional multinomial sampling distribution. Then from Example 16, we can express the multinomial density as

$$f(\mathbf{x}|\boldsymbol{\theta}) = \frac{m!}{\prod_{j=1}^{k} x_j!} \theta_k^m \exp\left(\sum_{j=1}^{k-1} x_j \log(\theta_j/\theta_k)\right)$$

and therefore, a conjugate prior distribution is

$$\begin{aligned} f(\boldsymbol{\theta}) &\propto (\theta_k{}^m)^a \exp\left(\sum_{j=1}^{k-1} b_j \log(\theta_j/\theta_k)\right) \quad \text{for} \quad \text{arbitrary } a, b_1, \ldots, b_{k-1} \\ &\propto \prod_{j=1}^{k-1} \theta_j{}^{b_j} \theta_k^{a-\sum_{j=1}^{k-1} b_j} \propto \prod_{j=1}^{k} \theta_j^{\alpha_j - 1} \end{aligned}$$

where $\alpha_j = b_j + 1$ for $j = 1, \ldots, k-1$ and $\alpha_k = a - \sum_{j=1}^{k} b_j + 1$.

The Dirichlet distribution

Definition 5

A random variable, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ is said to have a Dirichlet distribution, $\boldsymbol{\theta} \sim \mathcal{D}(\alpha_1, \ldots, \alpha_k)$ if

$$p(\boldsymbol{\theta}) = \frac{\Gamma\left(\sum_{j=1}^k \alpha_j\right)}{\prod_{j=1}^k \Gamma(\alpha_j)} \prod_{j=1}^k \theta_j^{\alpha_j - 1}$$

for $0 < \theta_j < 1$, $\sum_{j=1}^k \theta_j = 1$.

The Dirichlet distribution may be thought of as a generalization of the beta distribution. In particular, the marginal distribution of any θ_j is beta, that is $\theta_j \sim \mathcal{B}(\alpha_j, \alpha_0 - \alpha_j)$, where $\alpha_0 = \sum_{j=1}^k \alpha_j$.

Also, the moments of the Dirichlet distribution are easily evaluated:

$$
\begin{aligned}
E[\theta_j] &= \frac{\alpha_j}{\alpha_0} \\
V[\theta_j] &= \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)} \\
Cov[\theta_i \theta_j] &= -\frac{\alpha_i \alpha_j}{\alpha_0^2(\alpha_0 + 1)}
\end{aligned}
$$

Prediction and posterior distribution

Theorem 14

Let $\mathbf{X}|\boldsymbol{\theta} \sim \mathcal{MN}(m, \boldsymbol{\theta})$ and $\boldsymbol{\theta} \sim \mathcal{D}(\alpha_1, \ldots, \alpha_k)$. Then:

$$P(\mathbf{X} = \mathbf{x}) = \frac{m!\Gamma(\alpha_0)}{\Gamma(m + \alpha_0)} \prod_{j=1}^{k} \frac{\Gamma(x_j + \alpha_j)}{x_j!\Gamma(\alpha_j)} \quad \text{for } x_j \geq 0, \sum_{j=1}^{k} x_j = m$$

where $\alpha_0 = \sum_{j=1}^{k} \alpha_j$. Also

$$\begin{aligned} E[X_j] &= m\frac{\alpha_j}{\alpha_0} \\ V[X_j] &= m(\alpha_0 + m)\frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)} \\ Cov[X_i, X_j] &= -m(\alpha_0 + m)\frac{\alpha_i\alpha_j}{\alpha_0^2(\alpha_0 + 1)} \end{aligned}$$

Given a sample, $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of multinomial data, then

$$\boldsymbol{\theta}|\mathbf{x} \sim \mathcal{D}\left(\alpha_1 + \sum_{j=1}^{n} x_{1j}, \ldots, \alpha_k + \sum_{j=1}^{n} x_{kj}\right).$$

Proof Exercise ∎

Canonical form exponential family distributions

Suppose that we have a standard exponential family distribution

$$f(\mathbf{x}|\boldsymbol{\theta}) = C(\boldsymbol{\theta})h(\mathbf{x}) \exp\left(\sum_{j=1}^{k} \phi_j(\boldsymbol{\theta}) s_j(\mathbf{x})\right).$$

Define the transformed variables $Y_j = s_j(\mathbf{X})$ and transformed parameters $\phi_j = \phi_j(\boldsymbol{\theta})$ for $j = 1, \ldots k$. Then the density of $\mathbf{y}|\boldsymbol{\phi}$ is of *canonical form*.

Definition 6

The density

$$f(\mathbf{y}|\boldsymbol{\phi}) = D(\mathbf{y}) \exp\left(\sum_{j=1}^{k} y_j \phi_j - e(\boldsymbol{\phi})\right)$$

is called the canonical form representation of the exponential family distribution.

Conjugate analysis

Clearly, a conjugate prior for $\boldsymbol{\phi}$ takes the form

$$p(\boldsymbol{\phi}) \propto \exp\left(\sum_{j=1}^{k} b_j \phi_j - a e(\boldsymbol{\phi})\right) \propto \exp\left(\sum_{j=1}^{k} a B_j \phi_j - a e(\boldsymbol{\phi})\right)$$

where $B_j = \frac{b_j}{a}$ for $j = 1, \ldots, k$. We shall call this form the *canonical prior* for $\boldsymbol{\phi}$. If a sample of size n is observed, then

$$\begin{aligned} p(\boldsymbol{\phi}|\mathbf{y}) &\propto \exp\left(\sum_{j=1}^{k} \frac{a B_j + n\bar{y}_{\cdot j}}{a+n} \phi_j - (a+n) e(\boldsymbol{\phi})\right) \\ &\propto \exp\left(\frac{a\mathbf{B} + n\bar{\mathbf{y}}}{a+n}^T \boldsymbol{\phi} - (a+n) e(\boldsymbol{\phi})\right) \end{aligned}$$

The updating process for conjugate models involves a simple weighted average representation.

The posterior mean as a weighted average

Suppose that we observe a sample of n data from a canonical exponential family distribution as in Definition 6 and that we use a canonical prior distribution

$$p(\boldsymbol{\phi}) \propto \exp\left(\sum_{j=1}^{k} aB_j\phi_j - ae(\boldsymbol{\phi})\right) \propto \exp\left(a\mathbf{B}^T\boldsymbol{\phi} - ae(\boldsymbol{\phi})\right).$$

Then we can demonstrate the following theorem.

Theorem 15
$E\left[\nabla e(\boldsymbol{\phi})\right] = \frac{a\mathbf{B}+n\bar{\mathbf{y}}}{a+n}$ where $\left[\nabla e(\boldsymbol{\phi})\right]_j = \frac{\partial}{\partial\phi_j}e(\boldsymbol{\phi})$.

Proof As we have shown that the prior is conjugate, it is enough to prove that, a priori, $E[\nabla e(\boldsymbol{\phi})] = \mathbf{B}$. However,

$$a\left(\mathbf{B} - E[\nabla e(\boldsymbol{\phi})]\right) = \int a\left(\mathbf{B} - \nabla e(\boldsymbol{\phi})\right)p(\boldsymbol{\phi})\,d\boldsymbol{\phi} = \int \nabla p(\boldsymbol{\phi})\,d\boldsymbol{\phi}$$

∎

Example 17

Let $X|\theta \sim \mathcal{P}(\theta)$. Then, from Example 13, we can write

$$f(x|\theta) = e^{-\theta}\frac{1}{x!}\exp(x\log\theta) = \frac{1}{x!}\exp(x\log\theta - \theta) = \frac{1}{x!}\exp\left(x\phi - e^{\phi}\right),$$

where $\phi = e^{\theta}$, in canonical exponential family form. We already know that the gamma prior density $\theta \sim \mathcal{G}(\alpha, \beta)$ is conjugate here. Thus,

$$\begin{aligned} p(\theta) &\propto \theta^{\alpha-1}e^{-\beta\theta} \Rightarrow \\ p(\phi) &\propto (\log\phi)^{\alpha}\phi^{-\beta} \\ &\propto \exp\left(\beta\frac{\alpha}{\beta}\phi - \beta e^{\phi}\right) \end{aligned}$$

in canonical form. Now $\nabla\phi = \frac{d}{d\phi}e^{\phi} = e^{\phi} = \theta$. Thus, from Theorem 15, we have $E[\theta|\mathbf{x}] = \frac{\beta}{\beta+n}\frac{\alpha}{\beta} + \frac{n}{\alpha+n}\bar{x}$, a weighted average of the prior mean and the MLE.

Example 18

Consider the Bernoulli trial $f(x|\theta) = \theta^x(1-\theta)^{1-x}$ for $x = 0, 1$. We have

$$\begin{aligned} f(x|\theta) &= (1-\theta)\left(\frac{\theta}{1-\theta}\right)^x \\ &= \exp\left(x\log\frac{\theta}{1-\theta} - \log\frac{1}{1-\theta}\right) \\ &= \exp\left(x\phi - \log(1+e^{\phi})\right) \end{aligned}$$

in canonical form, where $\phi = \log\frac{\theta}{1-\theta}$.

Thus, the canonical prior for ϕ must take the form ,

$$\begin{aligned} p(\phi) &\propto \exp\left(aB\phi - a\log(1+e^{\phi})\right) \\ &\propto (1+e^{\phi})^{-a}\exp(aB\phi). \end{aligned}$$

This implies, using the standard change of variables formula, that

$$\begin{aligned} p(\theta) &\propto (1-\theta)^a \left(\frac{\theta}{1-\theta}\right)^{aB} \frac{1}{\theta(1-\theta)} \\ &\propto \theta^{aB-1}(1-\theta)^{a(1-B)-1} \end{aligned}$$

which we can recognize as a beta distribution, $\mathcal{B}(\alpha, \beta)$, with parameters $\alpha = aB$ and $\beta = a(1-B)$.

Now $\nabla \log(1+e^\phi) = \frac{e^\phi}{1+e^\phi} = \theta$ and so, from Theorem 15, given a sample of n Bernoulli trials, we have

$$E[\theta|\mathbf{x}] = \frac{aB + n\bar{x}}{a+n} = w\frac{\alpha}{\alpha+\beta} + (1-w)\bar{x}$$

where $w = \frac{\alpha+\beta}{\alpha+\beta+n}$.

We previously derived this formula by standard Bayesian analysis on page 95.

Mixtures of conjugate priors

Suppose that a simple conjugate prior distribution $p(\cdot) \in \mathcal{P}$ does not well represent our prior beliefs. An alternative is to consider a mixture of conjugate prior distributions

$$\sum_{i=1}^{k} w_i p_i(\boldsymbol{\theta})$$

where $0 < w_i < 1$ and $\sum_{i=1}^{k} w_i = 1$ and $p_i(\cdot) \in \mathcal{P}$.

Note that Dalal and Hall (1983) demonstrate that any prior density for an exponential family can be approximated arbitrarily closely by a mixture of conjugate distributions.

It is easy to demonstrate that given a conjugate prior mixture distribution the posterior distribution is also a mixture of conjugate densities.

Proof Using Bayes theorem, we have

$$\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{x}) &\propto f(\mathbf{x}|\boldsymbol{\theta}) \sum_{i=1}^{k} w_i p_i(\boldsymbol{\theta}) \\
&\propto \sum_{i=1}^{k} w_i p_i(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) \\
&\propto \sum_{i=1}^{k} w_i \left(\int p_i(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta} \right) \frac{p_i(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta})}{\int p_i(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta}} \\
&\propto \sum_{i=1}^{k} w_i \left(\int p_i(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta} \right) p_i(\boldsymbol{\theta}|\mathbf{x})
\end{aligned}$$

where $p_i(\cdot|\mathbf{x}) \in \mathcal{P}$ because it is assumed that $p_i(\cdot)$ is conjugate and therefore $p(\boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{k} w_i^\star p_i(\boldsymbol{\theta}|\mathbf{x})$ where $w_i^\star = \frac{w_i \left(\int p_i(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta} \right)}{\sum_{j=1}^{k} w_j \left(\int p_j(\boldsymbol{\theta}) f(\mathbf{x}|\boldsymbol{\theta}) d\boldsymbol{\theta} \right)}$. ∎

Example 19

In Example 11, suppose that we use a mixture prior distribution:

$$\theta \sim 0.25\mathcal{B}(1,1) + 0.75\mathcal{B}(5,5).$$

Then, the posterior distribution is given by

$$\begin{aligned}
p(\theta|\mathbf{x}) &\propto \left(0.25 \times 1 + 0.75 \times \frac{1}{B(5,5)}\theta^{5-1}(1-\theta)^{5-1}\right)\theta^9(1-\theta)^3 \\
&\propto 0.25\theta^{10-1}(1-\theta)^{4-1} + 0.75\frac{1}{B(5,5)}\theta^{14-1}(1-\theta)^{8-1} \\
&\propto B(10,4)\frac{1}{B(10,4)}\theta^{10-1}(1-\theta)^{4-1} + 3\frac{B(14,8)}{B(5,5)}\frac{1}{B(14,8)}\theta^{14-1}(1-\theta)^{8-1} \\
&= w^\star\frac{1}{B(10,4)}\theta^{10-1}(1-\theta)^{4-1} + (1-w^\star)\frac{1}{B(14,8)}\theta^{14-1}(1-\theta)^{8-1}
\end{aligned}$$

where $w^\star = \frac{B(10,4)}{B(10,4)+3B(14,8)/B(5,5)} = 0.2315$.

Thus, the posterior distribution is a mixture of $\mathcal{B}(10,4)$ and $\mathcal{B}(14,8)$ densities with weights 0.2315 and 0.7685 respectively.

Software for conjugate models

Some general software for undertaking conjugate Bayesian analysis has been developed.

- First Bayes is a slightly dated but fairly complete package
- The book on Bayesian computation by Albert (2009) gives a number of `R` routines for fitting conjugate and non conjugate models all contained in the `R` LearnBayes package.

References

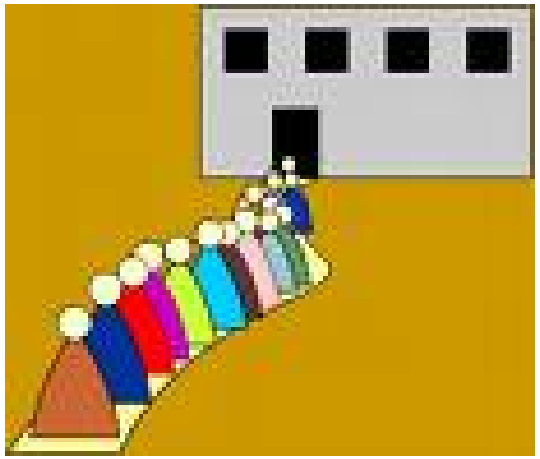
Albert, J. (2009) Bayesian Computation with R. Berlin: Springer.

Dalal, S.R. and Hall, W.J. (1983). Approximating priors by mixtures of natural conjugate priors. *Journal of the Royal Statistical Society Series B*, **45**, 278–286.

Haldane, J.B.S. (1931). A note on inverse probability. *Proceedings of the Cambridge Philosophical Society*, **28**, 55–61.

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. Cambridge MA: Harvard University Press.

Application I: Bayesian inference for Markovian queuing systems



A queue of children

We will study inference for the $M/M/1$ queueing system, as developed by Armero and Bayarri (1994a).

The $M/M/1$ queueing system

The $M(\lambda)/M(\mu)/1$ system is a queue with a single server and exponential inter-arrival and service times with rates λ and μ respectively.

Under this model, clients arrive at the checkout according to a Markov process with rate λ and if the server is empty, they start an exponential service time process with rate μ. If the server is busy, the client joins the queue of customers waiting to be served.

Stability and equilibrium distributions

In practical analysis, it is of particular importance to study the conditions under which such a queueing system is stable, i.e. when the queue length does not go off to infinity as more customers arrive. It can be shown that the system is stable if the *traffic intensity*, defined by $\rho = \lambda/\mu$ is strictly less than 1.

In the case that the system is stable, it is interesting to consider the equilibrium or stationary distribution of the queue size, client waiting times etc.

It can be shown that the equilibrium distribution of the number of clients in the system, N is geometric;

$$P(N = n|\rho) = (1-\rho)\rho^n \quad \text{for } N = 0, 1, 2, \ldots \quad \text{with mean} E[N|\rho] = \frac{\rho}{1-\rho}.$$

Also, the limiting distribution of the time, W, spent in the system by a customer is exponential; $W|\lambda, \mu \sim \mathcal{E}(\mu - \lambda)$, with mean $E[W|\lambda, \mu] = \frac{1}{\mu-\lambda}$.

The distributions of other variables of interest such as the duration of a busy period are also well known. See e.g. Gross and Harris (1985).

Experiment and inference

In real systems, the arrival and service rates will be unknown and we must use inferential techniques to estimate these parameters and the system characteristics. A first step is to find a reasonable way of collecting data.

The simplest experiment is that of observing n_l inter-arrival times and n_s service times. In this case, the likelihood is

$$l(\lambda, \mu | \mathbf{x}, \mathbf{y}) \propto \lambda^{n_l} e^{-\lambda t_l} \mu^{n_s} e^{-\mu t_s}$$

where t_l and t_s are the sums of inter-arrival and service times respectively.

If we use conjugate, independent gamma prior distributions,

$$\lambda \sim \mathcal{G}(\alpha_l, \beta_l) \quad \mu \sim \mathcal{G}(\alpha_s, \beta_s),$$

then the posterior distributions are also gamma distributions:

$$\lambda | \mathbf{x} \sim \mathcal{G}(\alpha_l + n_l, \beta_l + t_l) \quad \mu | \mathbf{y} \sim \mathcal{G}(\alpha_s + n_s, \beta_s + t_s).$$

Estimation of the traffic intensity

It is straightforward to estimate the mean of the traffic intensity ρ. We have

$$\begin{aligned} E[\rho|\mathbf{x},\mathbf{y}] &= E\left[\frac{\lambda}{\mu}\middle|\mathbf{x},\mathbf{y}\right] \\ &= E[\lambda|\mathbf{x}]E\left[\frac{1}{\mu}\middle|\mathbf{y}\right] \\ &= \frac{\alpha_l+n_l}{\beta_l+t_l}\frac{\beta_s+t_s}{\alpha_s+n_s-1} \end{aligned}$$

We can also evaluate the distribution of ρ by recalling that if $\phi \sim \mathcal{G}\left(\frac{a}{2},\frac{b}{2}\right)$, then $b\phi \sim \chi^2_a$ is chi-square distributed.

The probability that the system is stable

Thus, we have

$$2(\beta_l + t_l)\lambda|\mathbf{x} \sim \chi^2_{2(\alpha_l+n_l)} \quad 2(\beta_s + t_s)\mu|\mathbf{y} \sim \chi^2_{2(\alpha_s+n_s)}$$

and therefore, recalling that the ratio of two independent χ^2 variables divided by their degrees of freedom is an F distributed variable, we have

$$\frac{(\beta_l + t_l)(\alpha_s + n_s)}{(\alpha_l + n_l)(\beta_s + t_s)}\rho|\mathbf{x}, \mathbf{y} \sim \mathcal{F}^{2(\alpha_l+n_l)}_{2(\alpha_s+n_s)}.$$

The posterior probability that the system is stable is

$$\begin{aligned} p &= P(\rho < 1|\mathbf{x}, \mathbf{y}) \\ &= P\left(F < \frac{(\beta_l + t_l)(\alpha_s + n_s)}{(\alpha_l + n_l)(\beta_s + t_s)}\right) \quad \text{where } F \sim \mathcal{F}^{2(\alpha_l+n_l)}_{2(\alpha_s+n_s)} \end{aligned}$$

which can be easily evaluated in e.g. `Matlab` or `R`.

Estimation of the queue size in equilibrium

If p is large, it is natural to suppose that the system is stable. In this case, it is interesting to carry out prediction for the equilibrium distribution of the number of clients in the system. We have

$$\begin{aligned} P(N = n|\mathbf{x}, \mathbf{y}, \text{equilibrium}) &= P(N = n|\mathbf{x}, \mathbf{y}, \rho < 1) \\ &= \int_0^1 (1-\rho)\rho^n p(\rho|\mathbf{x}, \mathbf{y}, \rho < 1)\, d\rho \\ &= \frac{1}{p}\int_0^1 (1-\rho)\rho^n p(\rho|\mathbf{x}, \mathbf{y})\, d\rho \end{aligned}$$

We will typically need numerical methods to evaluate this integral. One possibility is to use simple *numerical integration*. Another technique is to use *Monte Carlo sampling*.

Aside: Monte Carlo sampling

Suppose that we wish to estimate an integral

$$E[g(X)] = \int g(x)f(x)\,dx$$

where X is a random variable with density $f(\cdot)$. Then, if we draw a sample, say $x_1, \ldots, x_M$ of size from $f(\cdot)$ then, under certain regularity conditions, as $M \to \infty$, we have

$$\bar{g} = \frac{1}{M}\sum_{i=1}^{M} g(x_i) \to E[g(X)].$$

In order to assess the precision of a Monte Carlo based estimator, a simple technique is to calculate the confidence band $\bar{g} \pm 2s_g/\sqrt{M}$ where $s_g^2 = \frac{1}{M}\sum_{i=1}^{M}\left(g(x_i) - \bar{g}\right)^2$ is the sample variance of $g(\cdot)$. The Monte Carlo sample size can be increased until the size of the confidence band goes below a certain preset precision ϵ.

We will discuss Monte Carlo methods in more detail in chapter 6.

Using Monte Carlo to estimate the system size distribution

We can set up a generic algorithm to generate a Monte Carlo sample.

1. Fix a large value M.

2. For $i = 1, \ldots, M$:

   (a) Generate $\lambda_i \sim \mathcal{G}(\alpha_l + n_l, \beta_l + n_l)$ and $\mu_i \sim \mathcal{G}(\alpha_s + n_s, \beta_s + n_s)$.
   (b) Set $\rho_i = \lambda_i / \mu_i$.
   (c) If $\rho_i > 1$, go to a).

Given the Monte Carlo sampled data, we can then estimate the queue size probabilities, $P(N = n | \mathbf{x}, \mathbf{y}, \text{equilibrium}) \approx \frac{1}{M} \sum_{i=1}^{M} (1 - \rho_i) \rho_i^n$ for $n = 0, 1, 2, \ldots$ and the waiting time distribution, $P(W \leq w | \mathbf{x}, \mathbf{y}, \text{equilibrium}) \approx 1 - \frac{1}{M} \sum_{i=1}^{M} e^{-(\mu_i - \lambda_i)}$.

Estimating the mean of the equilibrium system size distribution

It is easier to evaluate the predictive moments of N. We have

$$\begin{aligned} E[N|\mathbf{x},\mathbf{y},\text{equilibrium}] &= E[E[N|\rho]|\mathbf{x},\mathbf{y},\rho<1] \\ &= E\left[\frac{\rho}{1-\rho}\middle|\mathbf{x},\mathbf{y},\rho<1\right] \\ &= \frac{1}{p}\int_0^1 \frac{\rho}{1-\rho}p(\rho|\mathbf{x},\mathbf{y})\,d\rho = \infty \end{aligned}$$

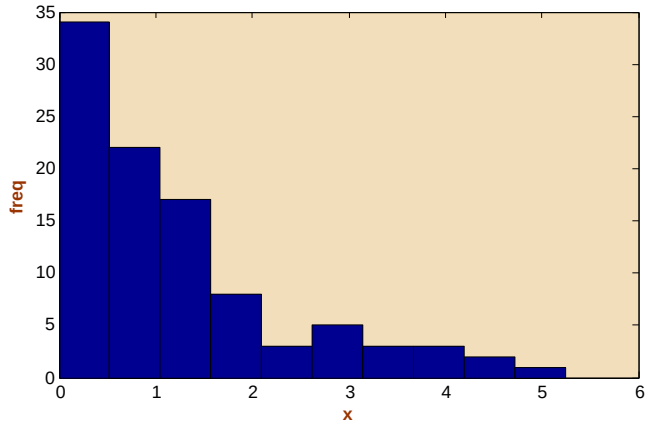
and thus, these moments do not exist.

This is a general characteristic of inference in queueing systems with Markovian inter-arrival or service processes. The same thing also happens with the equilibrium waiting time or busy period distributions. See Armero and Bayarri (1994a) or Wiper (1998).

The problem stems from the use of (independent) prior distributions for λ and μ with $p(\lambda=\mu)>0$ and can be partially resolved by assuming *a priori* that $P(\lambda\geq\mu)=0$. See Armero and Bayarri (1994b) or Ruggeri et al (1996).

Example

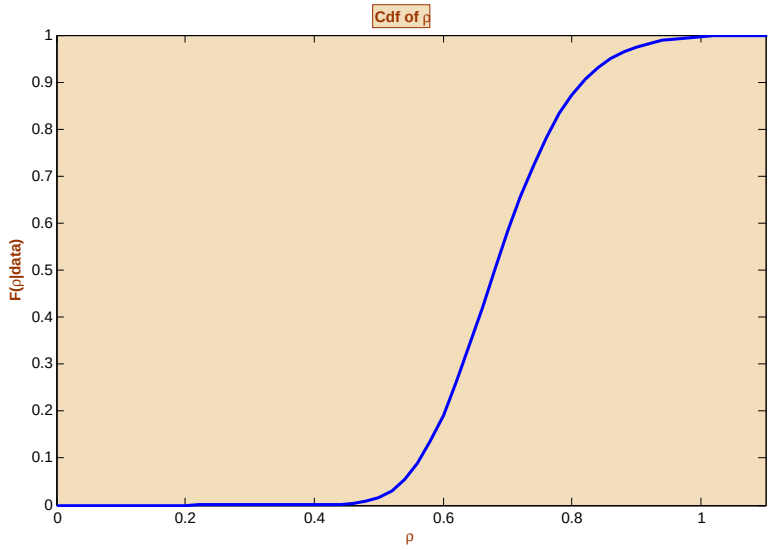
Hall (1991) gives collected inter-arrival and service time data for 98 users of an automatic teller machine in Berkeley, USA. We shall assume here that inter-arrival and service times can be modeled as exponential variables.



The sufficient statistics are $n_a = n_s = 98$, $t_a = 119.71$ and $t_s = 81.35$ minutes respectively. We will assume the $improper$ prior distributions $p(\lambda) \propto \frac{1}{\lambda}$ and $p(\mu) \propto \frac{1}{\mu}$ which we have seen earlier as limiting cases of the conjugate gamma priors. Then $\lambda|\mathbf{x} \sim \mathcal{G}(98, 119.71)$ and $\mu|\mathbf{y} \sim \mathcal{G}(98, 119.71)$.

The posterior distribution of the traffic intensity

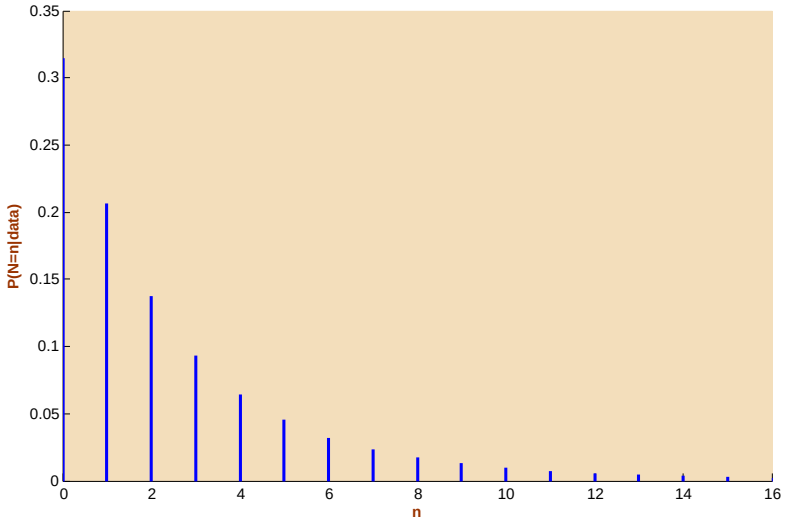
The posterior expected value of ρ is $E[\rho|\mathbf{x},\mathbf{y}] \approx 0.69$ and the distribution function $F(\rho|\mathbf{x},\mathbf{y})$ is illustrated below.



The posterior probability that the system is stable is 0.997.

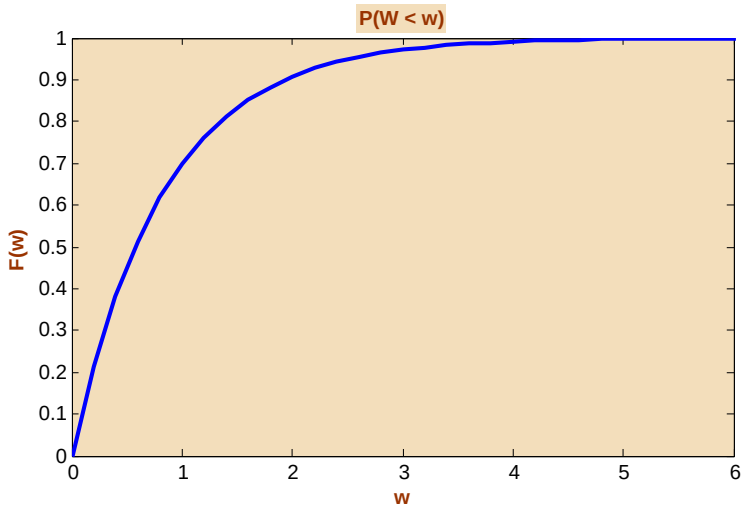
The posterior distribution of N

We used a large Monte Carlo sample to estimate the posterior density of N. There is only a very small probability that there are more than 15 clients in the system.



The posterior distribution of W

There is only a very small probability that a client spends over 5 minutes in the system.



Extensions

- Calculation of busy period and transient time distributions.

- Extension to other queueing systems:

  ◇ Markovian systems with various or unlimited numbers of servers, e.g. Armero and Bayarri (1997).
  ◇ Networks of queues. Armero and Bayarri (1999).
  ◇ Non Markovian systems. See e.g. Wiper (1998), Ausín et al (2004).

References

Armero, C. and Bayarri, M.J. (1994a). Bayesian prediction in M/M/1 queues. *Queueing Systems*, **15**, 419-426.

Armero, C. and Bayarri, M.J. (1994b). Prior assessments for prediction in queues. *The Statistician*, **43**, 139-153

Armero, C. and Bayarri, M.J. (1997). Bayesian analysis of a queueing system with unlimited service. *Journal of Statistical Planning and Inference*, **58**, 241–261.

Armero, C. and Bayarri, M.J. (1999). Dealing with Uncertainties in Queues and Networks of Queues: A Bayesian Approach. In: *Multivariate Analysis, Design of Experiments and Survey Sampling*, Ed. S. Ghosh. New York: Marcel Dekker, pp 579–608.

Ausín, M.C., Wiper, M.P. and Lillo, R.E. (2004). Bayesian estimation for the M/G/1 queue using a phase type approximation. *Journal of Statistical Planning and Inference*, **118**, 83–101.

Gross, D. and Harris, C.M. (1985). *Fundamentals of Queueing Theory* (2nd ed.). New York: Wiley.

Hall, R.W. (1991). *Queueing Methods*. New Jersey: Prentice Hall.

Ruggeri, F., Wiper, M.P. and Rios Insua, D. (1996). Bayesian models for correlation in M/M/1 queues. *Quaderno IAMI 96.8*, CNR-IAMI, Milano.

Wiper, M.P. (1998). Bayesian analysis of Er/M/1 and Er/M/c queues. *Journal of Statistical Planning and Inference*, **69**, 65–79.