Diffuse Priors in the LOS Case Study

In specifying a diffuse prior for \( \lambda \) in the LOS case study, several alternatives to \( \Gamma(\epsilon, \epsilon) \) were suggested in class, including \( \Gamma(1, \epsilon), \Gamma(\alpha, \beta) \) for some large \( \alpha \) (like 20, to get a roughly normal prior) and small \( \beta \) (like 1, to have a small prior sample size), and \( U(0, C) \) for some cutoff \( C \) (like 4) chosen to avoid truncation of the likelihood function, where \( U(a, b) \) denotes the uniform distribution on \((a, b)\).

\[
> \text{plot( p( lambda, 0.001, 0.001 ), lambda = 0 .. 4, v = 0 .. 0.05, color = black );}
\]

\[
> \text{plot( p( lambda, 1.0, 0.001 ), lambda = 0 .. 4, color = black );}
\]
Diffuse Priors (continued)

\Gamma(1, \epsilon) doesn't look promising initially as a flat prior, but that's a consequence of Maple's default choice of vertical axis.

\begin{verbatim}
> plot( p( lambda, 1.0, 0.001 ), lambda = 0 .. 4, v = 0 .. 0.05, color = black );
\end{verbatim}

\begin{verbatim}
> plot( p( lambda, 20, 1 ), lambda = 0 .. 4, color = black );
\end{verbatim}
Diffuse Priors (continued)

> plot( p( lambda, 20, 1 ), lambda = 0 .. 40, color = black );

Γ(20,1) does indeed look not far from Gaussian, and at first it may appear that it is indeed relatively flat in the region where the likelihood is appreciable (λ ∈ (1.0,3.3)), but we'll see below that it's actually rather more informative than we intend.

Recalling that the mean and SD of a Γ(α, β) random quantity are \( \frac{α}{β} \) and \( \sqrt{\frac{α}{β^3}} \), respectively, and that when used as a prior with the Poisson likelihood the Γ(α, β) distribution acts like a dataset with prior sample size β, you can construct the following table:

<table>
<thead>
<tr>
<th>Prior</th>
<th>( \alpha = \frac{C}{2} )</th>
<th>( \beta = \frac{C}{\sqrt{12}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( U(0,C) ) for ( C &gt; 4 )</td>
<td>30</td>
<td>14</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>Sample Size</td>
<td>Mean</td>
</tr>
<tr>
<td>0.001</td>
<td>0.001</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0.001</td>
<td>1000</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>20</td>
<td>0.001</td>
<td>20000</td>
</tr>
</tbody>
</table>
Diffuse Priors (continued)

The $\Gamma(1, \epsilon)$ prior leads to an analysis that’s essentially equivalent to the integrated likelihood (fiducial) approach back on p. 68, and the $U(0, C)$ prior for $C > 4$ (say) produces similar results: $U(0, C)$ yields the $\Gamma(s + 1, n)$ posterior truncated to the right of $C$ (and this truncation has no effect if you choose $C$ big enough).

You might say that the $U(0, C)$ distribution has a prior sample size of 0 in this analysis, and its prior mean $\frac{C}{2}$ and SD $\frac{C}{\sqrt{12}}$ (both of which can be made arbitrarily large by letting $C$ grow without bound) are irrelevant (an example of how intuition can change when you depart from the class of conjugate priors).

```r
> plot( { p( lambda, 29.001, 14.001 ), p( lambda, 30, 14.001 ),
        p( lambda, 49, 15 ), p( lambda, 49, 14.001 ) }, lambda = 0 .. 6,
        color = black );
```

The moral seems to be that with only $n = 14$ observations, some care is needed to achieve a prior that doesn't affect the posterior very much, if that's your goal.
Gaussian Modeling (continued)

- The **posterior**, considered as an **information source**, is Gaussian, and the posterior mean is a **weighted average** of the prior mean and data mean, with weights given by the **prior** and **data precisions**;

- The **posterior precision** (the reciprocal of the posterior variance) is just the **sum** of the prior and data precisions (this is why Bayesians invented the idea of precision—on this scale **knowledge** about $\mu$ in model (74) is **additive**); and

  - **Rewriting** $\mu_*$ as
    \[
    \mu_* = \frac{\left(\frac{1}{\sigma_\mu^2}\right) \mu_0 + \left(\frac{n}{\sigma_0^2}\right) \bar{y}}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}} = \frac{\left(\frac{\sigma_0^2}{\sigma_\mu^2}\right) \mu_0 + n\bar{y}}{\frac{\sigma_0^2}{\sigma_\mu^2} + n},
    \]
    (81)

  you can see that the **prior sample size** is
    \[
    n_0 = \frac{\sigma_0^2}{\sigma_\mu^2} = \frac{1}{\left(\frac{\sigma_\mu}{\sigma_0}\right)^2},
    \]
    (82)

  which makes sense: the **bigger** $\sigma_\mu$ is in relation to $\sigma_0$, the **less prior information** is being incorporated in the conjugate updating (79).

**Bayesian inference with multivariate $\theta$.** Returning now to (72) with $\sigma^2$ unknown, (as mentioned above) this model has a ($p = 2$)-dimensional **parameter vector** $\theta = (\mu, \sigma^2)$.

When $p > 1$ you can still use Bayes' Theorem directly to obtain the **joint posterior distribution**, \[
    p(\theta|y) = p(\mu, \sigma^2|y) = cp(\theta) l(\theta|y)
    = cp(\mu, \sigma^2) l(\mu, \sigma^2|y),
    \]
(83)
Multivariate Unknown $\theta$

where $y = (y_1, \ldots, y_n)$, although making this calculation directly requires a $p$-dimensional integration to evaluate the normalizing constant $c$; for example, in this case

$$
c = [p(y)]^{-1} = \left( \int \int p(\mu, \sigma^2, y) \, d\mu \, d\sigma^2 \right)^{-1}
= \left( \int \int p(\mu, \sigma^2) l(\mu, \sigma^2 | y) \, d\mu \, d\sigma^2 \right)^{-1}.
$$

(84)

Usually, however, you’ll be more interested in the marginal posterior distributions, in this case $p(\mu | y)$ and $p(\sigma^2 | y)$.

Obtaining these requires $p$ integrations, each of dimension $(p - 1)$, a process that people refer to as marginalization or integrating out the nuisance parameters—for example,

$$
p(\mu | y) = \int_0^\infty p(\mu, \sigma^2 | y) \, d\sigma^2.
$$

(85)

Predictive distributions also involve a $p$-dimensional integration: for example, with $y = (y_1, \ldots, y_n)$,

$$
p(y_{n+1} | y) = \int \int p(y_{n+1}, \mu, \sigma^2 | y) \, d\mu \, d\sigma^2
= \int \int p(y_{n+1} | \mu, \sigma^2) p(\mu, \sigma^2 | y) \, d\mu \, d\sigma^2.
$$

(86)

And, finally, if you’re interested in a function of the parameters, you have some more hard integrations ahead of you.

For instance, suppose you wanted the posterior distribution for the coefficient of variation $\lambda = g_1(\mu, \sigma^2) = \frac{\mu}{\sqrt{\sigma^2}}$ in model (72).
Multivariate Unknown $\theta$

Then one fairly direct way to get this posterior (e.g., Bernardo and Smith, 1994) is to (a) introduce a second function of the parameters, say $\eta = g_2(\mu, \sigma^2)$, such that the mapping $f = (g_1, g_2)$ from $(\mu, \sigma^2)$ to $(\lambda, \eta)$ is invertible; (b) compute the joint posterior for $(\lambda, \eta)$ through the usual change-of-variables formula

$$p(\lambda, \eta|y) = p_{\mu,\sigma^2}[f^{-1}(\lambda, \eta)|y] |J_{f^{-1}}(\lambda, \eta)|,$$

where $p_{\mu,\sigma^2}(\cdot, \cdot|y)$ is the joint posterior for $\mu$ and $\sigma^2$ and $|J_{f^{-1}}|$ is the determinant of the Jacobian of the inverse transformation; and (c) marginalize in $\lambda$ by integrating out $\eta$ in $p(\lambda, \eta|y)$, in a manner analogous to (85).

Here, for instance, $\eta = g_2(\mu, \sigma^2) = \sqrt{\sigma^2}$ would create an invertible $f$, with inverse defined by $(\mu = \lambda\eta, \sigma^2 = \eta^2)$; the Jacobian determinant comes out $2\lambda\eta$ and (87) becomes

$$p(\lambda, \eta|y) = 2\lambda\eta p_{\mu,\sigma^2}(\lambda\eta, \eta^2|y).$$

This process involves two integrations, one (of dimension $p$) to get the normalizing constant that defines (87) and one (of dimension $(p - 1)$) to get rid of $\eta$.

You can see that when $p$ is a lot bigger than 2 all these integrals may create severe computational problems—this has been the big stumbling block for applied Bayesian work for a long time.

More than 200 years ago Laplace (1774)—perhaps the second applied Bayesian in history (after Bayes himself)—developed, as one avenue of solution to this problem, what people now call Laplace approximations to high-dimensional integrals of the type arising in Bayesian calculations (see, e.g., Tierney and Kadane, 1986).

Starting in the next case study after this one, we'll use another, computationally intensive, simulation-based approach: Markov chain Monte Carlo (MCMC).
Gaussian Modeling

Back to model (72). The conjugate prior for $\theta = (\mu, \sigma^2)$ in this model (e.g., Gelman et al., 1995) turns out to be most simply described hierarchically:

$$\begin{align*}
\sigma^2 &\sim \text{SI-}\chi^2(\nu_0, \sigma_0^2) \\
(\mu|\sigma^2) &\sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right). \tag{88}
\end{align*}$$

Here saying that $\sigma^2 \sim \text{SI-}\chi^2(\nu_0, \sigma_0^2)$, where SI stands for scaled inverse, amounts to saying that the precision $\tau = \frac{1}{\sigma^2}$ follows a scaled $\chi^2$ distribution with parameters $\nu_0$ and $\sigma_0^2$.

The scaling is chosen so that $\sigma_0^2$ can be interpreted as a prior estimate of $\sigma^2$, with $\nu_0$ the prior sample size of this estimate (i.e., think of a prior data set with $\nu_0$ observations and sample SD $\sigma_0$).

Since $\chi^2$ is a special case of the Gamma distribution, SI-$\chi^2$ must be a special case of the inverse Gamma family—its density (see Gelman et al. (1995), Appendix A) is

$$\begin{align*}
\sigma^2 &\sim \text{SI-}\chi^2(\nu_0, \sigma_0^2) \leftrightarrow \\
p(\sigma^2) &= \frac{(\frac{1}{2}\nu_0)^{\frac{1}{2}\nu_0}}{\Gamma(\frac{1}{2}\nu_0)} (\sigma_0^2)^{\frac{1}{2}\nu_0} (\sigma^2)^{-\left(1+\frac{1}{2}\nu_0\right)} \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right). \tag{89}
\end{align*}$$

As may be verified with Maple, this distribution has mean (provided that $\nu_0 > 2$) and variance (provided that $\nu_0 > 4$) given by

$$\begin{align*}
E(\sigma^2) &= \frac{\nu_0}{\nu_0 - 2} \sigma_0^2 \quad \text{and} \quad V(\sigma^2) = \frac{2\nu_0^2}{(\nu_0 - 2)^2(\nu_0 - 4)} \sigma_0^4. \tag{90}
\end{align*}$$
Gaussian Modeling (continued)

The parameters $\mu_0$ and $\kappa_0$ in the second level of the prior model (88), $(\mu|\sigma^2) \sim N(\mu_0, \frac{\sigma^2}{\kappa_0})$, have simple parallel interpretations to those of $\sigma_0^2$ and $\nu_0$: $\mu_0$ is the prior estimate of $\mu$, and $\kappa_0$ is the prior effective sample size of this estimate.

The likelihood function in model (72), with both $\mu$ and $\sigma^2$ unknown, is

$$l(\mu, \sigma^2|y) = c \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma^2} \exp \left[ -\frac{1}{2\sigma^2} (y_i - \mu)^2 \right]$$

$$= c (\sigma^2)^{-\frac{1}{2}n} \exp \left[ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right]$$

$$= c (\sigma^2)^{-\frac{1}{2}n} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2 \right) \right].$$

The expression in brackets in the last line of (91) is

$$[ \cdot ] = -\frac{1}{2\sigma^2} \left[ \sum_{i=1}^{n} y_i^2 + n(\mu - \bar{y})^2 - n\bar{y}^2 \right]$$

$$= -\frac{1}{2\sigma^2} \left[ n(\mu - \bar{y})^2 + (n-1)s^2 \right],$$

where $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the sample variance. Thus

$$l(\mu, \sigma^2|y) = c (\sigma^2)^{-\frac{1}{2}n} \exp \left\{ -\frac{1}{2\sigma^2} \left[ n(\mu - \bar{y})^2 + (n-1)s^2 \right] \right\},$$

and it's clear that the vector $(\bar{y}, s^2)$ is sufficient for $\theta = (\mu, \sigma^2)$ in this model, i.e., $l(\mu, \sigma^2|y) = l(\mu, \sigma^2|\bar{y}, s^2)$. 

94
Gaussian Analysis

Maple can be used to make 3D and contour plots of this likelihood function with the NB10 data:

\[
\begin{align*}
  l & := (\mu, \sigma_2, \bar{y}, s_2, n) \rightarrow \sigma_2^2 (-n/2) \times \\
  & \quad \times \exp\left(-n\left(\mu - \bar{y}\right)^2 + (n-1)s_2\right)/(2\sigma_2) ; \\
  l & := (\mu, \sigma_2, \bar{y}, s_2, n) \rightarrow \\
  & \frac{2}{\sigma_2} \left(-\frac{1}{2} n\right) \times \exp\left(-\frac{1}{2} \frac{n(\mu - \bar{y}) + (n-1)s_2}{\sigma_2}\right) \\
  & \sigma_2
\end{align*}
\]

\[
> \text{plotsetup( x11 );}
\]

\[
> \text{plot3d}( l(\mu, \sigma_2, 404.6, 42.25, 100 ), \mu = 402.6 .. 406.6, \\
  \sigma_2 = 25 .. 70 );
\]

You can use the mouse to rotate 3D plots and get other useful views of them:
Gaussian Analysis

The projection or shadow plot of $\mu$ looks a lot like a normal (or maybe a $t$) distribution.

And the shadow plot of $\sigma^2$ looks a lot like a Gamma (or maybe an inverse Gamma) distribution.
The contour plot shows that \( \mu \) and \( \sigma^2 \) are uncorrelated in the likelihood distribution, and the skewness of the marginal distribution of \( \sigma^2 \) is also evident.

**Posterior analysis.** Having adopted the conjugate prior (88), what I'd like next is simple expressions for the marginal posterior distributions \( p(\mu|y) \) and \( p(\sigma^2|y) \) and for predictive distributions like \( p(y_{n+1}|y) \).

Fortunately, in model (72) all of the integrations (such as (85) and (86)) may be done analytically (see, e.g., Bernardo and Smith 1994), yielding the following results:

\[
\begin{align*}
(\sigma^2|y, \mathcal{G}) & \sim \text{SI-}\chi^2(\nu_n, \sigma_n^2), \\
(\mu|y, \mathcal{G}) & \sim t_{\nu_n} \left( \mu_n, \frac{\sigma_n^2}{\kappa_n} \right), \quad \text{and} \\
(y_{n+1}|y, \mathcal{G}) & \sim t_{\nu_n} \left( \mu_n, \frac{\kappa_n + 1}{\kappa_n} \frac{\sigma_n^2}{\kappa_n} \right).
\end{align*}
\]
NB10 Gaussian Analysis

In the above expressions

\[ \nu_n = \nu_0 + n, \]
\[ \sigma_n^2 = \frac{1}{\nu_n} \left[ \nu_0 \sigma_0^2 + (n - 1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n} (\bar{y} - \mu_0)^2 \right], \quad (94) \]
\[ \mu_n = \frac{\kappa_0}{\kappa_0 + n} \mu_0 + \frac{n}{\kappa_0 + n} \bar{y}, \quad \text{and} \]
\[ \kappa_n = \kappa_0 + n, \]

\( \bar{y} \) and \( s^2 \) are the usual sample mean and variance of \( y \), and \( \mathcal{G} \) denotes the assumption of the Gaussian model.

Here \( t_\nu(\mu, \sigma^2) \) is a scaled version of the usual \( t_\nu \) distribution, i.e., \( W \sim t_\nu(\mu, \sigma^2) \iff \frac{W - \mu}{\sigma} \sim t_\nu. \)

The scaled \( t \) distribution (see, e.g., Gelman et al. (1995) Appendix A) has density

\[ \eta \sim t_\nu(\mu, \sigma^2) \iff p(\eta) = \frac{\Gamma \left[ \frac{1}{2}(\nu + 1) \right]}{\Gamma \left( \frac{1}{2} \nu \right) \sqrt{\nu \pi \sigma^2}} \left[ 1 + \frac{1}{\nu \sigma^2} (\eta - \mu)^2 \right]^{-\frac{1}{2}(\nu+1)}. \quad (95) \]

This distribution has mean \( \mu \) (as long as \( \nu > 1 \)) and variance \( \frac{\nu}{\nu - 2} \sigma^2 \) (as long as \( \nu > 2 \)).

Notice that, as with all previous conjugate examples, the posterior mean is again a weighted average of the prior mean and data mean, with weights determined by the prior sample size and the data sample size:

\[ \mu_n = \frac{\kappa_0}{\kappa_0 + n} \mu_0 + \frac{n}{\kappa_0 + n} \bar{y}. \quad (96) \]
NB10 Gaussian Analysis (continued)

**NB10 Gaussian Analysis.** Question (a): I don’t know anything about what NB10 is supposed to weigh (down to the nearest microgram) or about the accuracy of the NBS’s measurement process, so I want to use a diffuse prior for \( \mu \) and \( \sigma^2 \).

Considering the meaning of the hyperparameters, to provide little prior information I want to choose both \( \nu_0 \) and \( \kappa_0 \) close to 0.

Making them exactly 0 would produce an improper prior distribution (which doesn’t integrate to 1), but choosing positive values as close to 0 as you like yields a proper and highly diffuse prior.

You can see from (93, 94) that the result is then

\[
(\mu|y, \mathcal{G}) \sim t_n \left[ \bar{y}, \frac{(n-1)s^2}{n^2} \right] \equiv N \left( \frac{\bar{y}}{\frac{s^2}{n}} \right),
\]

i.e., with diffuse prior information (as with the Bernoulli model in the AMI case study) the 95% central Bayesian interval virtually coincides with the usual frequentist 95% confidence interval

\[
\bar{y} \pm t_{n-1}^{.975} \frac{s}{\sqrt{n}} = 404.6 \pm (1.98)(0.647) = (403.3, 405.9).
\]

Thus both \{frequentists who assume \( \mathcal{G} \)\} and \{Bayesians who assume \( \mathcal{G} \) with a diffuse prior\} conclude that NB10 weighs about 404.6\( \mu \)g below 10g, give or take about 0.65\( \mu \)g.

Question (b). If interest focuses on whether NB10 weighs less than some value like 405.25, when reasoning in a Bayesian way you can answer this question directly: the posterior distribution for \( \mu \) is shown below, and

\[
P_{B}(\mu < 405.25|y, \mathcal{G}, \text{diffuse prior}) = .85, \text{ i.e., your betting odds in favor of the proposition that } \mu < 405.25 \text{ are about 5.5 to 1.}
\]
When reasoning in a frequentist way \( P_F(\mu < 405.25) \) is **undefined**; about the best you can do is to test \( H_0: \mu < 405.25 \), for which the \( p \)-value would (approximately) be \( p = P_{F, \mu=405.25}(\bar{y} > 405.59) = 1 - .85 = .15 \), i.e., **insufficient evidence to reject** \( H_0 \) at the usual significance levels (note the **connection** between the \( p \)-value and the posterior probability, which arises in this example because the null hypothesis is **one-sided**).

**NB** The significance test tries to answer a **different question**: in Bayesian language it looks at \( P(\bar{y}|\mu) \) instead of \( P(\mu|\bar{y}) \).

Many people find the latter quantity **more interpretable**.

**Question (c).** We saw earlier that **in this model**

\[
(y_{n+1}|y, \mathcal{G}) \sim t_{\nu_n} \left[ \mu_n, \frac{\kappa_n + 1}{\kappa_n} \sigma_n^2 \right],
\]

and for \( n \) large and \( \nu_0 \) and \( \kappa_0 \) close to 0 this is \( (y_{n+1}|y, \mathcal{G}) \sim N(\bar{y}, s^2) \), i.e., a **95% posterior predictive interval** for \( y_{n+1} \) is (392, 418).
Model Expansion

A standardized version of this predictive distribution is plotted below, with the standardized NB10 data values superimposed.

![Graph showing standardized NB10 values](image)

It's evident from this plot (and also from the normal qqplot given earlier) that the Gaussian model provides a poor fit for these data—the three most extreme points in the data set in standard units are $-4.6$, $2.8$, and $5.0$.

With the symmetric heavy tails indicated in these plots, in fact, the empirical CDF looks quite a bit like that of a $t$ distribution with a rather small number of degrees of freedom.

This suggests revising the previous model by expanding it: embedding the Gaussian in the $t$ family and adding a parameter $k$ for tail-weight.

Unfortunately there's no standard closed-form conjugate choice for the prior on $k$.

A more flexible approach to computing is evidently needed—this is where Markov chain Monte Carlo methods (our next main topic) come in.
2.9 The Exponential Family

It was noticed a long time ago that many of the standard sampling distributions that you're likely to want to use in constructing likelihood functions have the **same general form**, which is referred to as the **exponential family**:

**Definition** (e.g., Bernardo and Smith, 1994):
Given data $y_1$ (a sample of size 1) and a parameter vector $\theta = (\theta_1, \ldots, \theta_k)$, the (marginal) sampling distribution $p(y_1|\theta)$ belongs to the **$k$-dimensional exponential family** if it can be expressed in the form

\[
p(y_1|\theta) = c f_1(y_1) g_1(\theta) \exp \left[ \sum_{j=1}^{k} \phi_j(\theta) h_j(y_1) \right] \quad (99)
\]

for $y_1 \in \mathcal{Y}$ and 0 otherwise; if $\mathcal{Y}$ does not depend on $\theta$ the family is called **regular**.

In this case the **joint distribution** $p(y|\theta)$ of a sample $y = (y_1, \ldots, y_n)$ of size $n$ which is conditionally IID from (99) (which also defines, as usual, the **likelihood function** $l(\theta|y)$) will be

\[
p(y|\theta) = l(\theta|y) = \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} f_1(y_i) [g_1(\theta)]^n \exp \left[ \sum_{j=1}^{k} \phi_j(\theta) \sum_{i=1}^{n} h_j(y_i) \right] \quad (100)
\]
The Exponential Family (continued)

This leads to another way to define the exponential family: in (99) take \( f(y) = \prod_{i=1}^{n} f_1(y_i) \)
and \( g(\theta) = [g_1(\theta)]^n \) to yield

**Definition**: Given data \( y = (y_1, \ldots, y_n) \) (a conditionally IID sample of size \( n \)) and a parameter vector \( \theta = (\theta_1, \ldots, \theta_k) \), the (joint) sampling distribution \( p(y|\theta) \) belongs to the \( k \)-dimensional exponential family if it can be expressed in the form

\[
p(y|\theta) = c f(y) g(\theta) \exp \left[ \sum_{j=1}^{k} \phi_j(\theta) \sum_{i=1}^{n} h_j(y_i) \right].
\]

(101)

Either way you can see that \( \{\sum_{i=1}^{n} h_1(y_i), \ldots, \sum_{i=1}^{n} h_k(y_i)\} \) is a set of sufficient statistics for \( \theta \) under this sampling model, because the likelihood \( l(\theta|y) \) depends on \( y \) only through the values of \( \{h_1, \ldots, h_k\} \).

I bring up the exponential family in part because, if the likelihood \( l(\theta|y) \) is of the form (101), then in searching for a conjugate prior \( p(\theta) \)—that is, a prior of the same functional form as the likelihood—you can see directly what will work:
The Exponential Family (continued)

\[ p(\theta) = c g(\theta)^{\tau_0} \exp \left[ \sum_{j=1}^{k} \phi_j(\theta) \tau_j \right], \tag{102} \]

for some \( \tau = (\tau_0, \ldots, \tau_k) \).

With this choice the \textbf{posterior} for \( \theta \) will be

\[ p(\theta|y) = c g(\theta)^{1+\tau_0} \exp \left\{ \sum_{j=1}^{k} \phi_j(\theta) \left[ \tau_j + \sum_{i=1}^{n} h_j(y_i) \right] \right\}, \tag{103} \]

which is indeed of the \textbf{same form} (in \( \theta \)) as (102).

As a first example, with \( s = \sum_{i=1}^{n} y_i \), the \textbf{Bernoulli/binomial} likelihood in (41) can be written

\[ l(\theta|y) = \theta^s (1-\theta)^{n-s} \]

\[ = (1-\theta)^n \left( \frac{\theta}{1-\theta} \right)^s \]

\[ = (1-\theta)^n \exp \left[ s \log \left( \frac{\theta}{1-\theta} \right) \right], \tag{104} \]

which shows (a) that this sampling distribution is a member of the \textbf{exponential family} with \( k = 1, \ g(\theta) = (1-\theta)^n, \ \phi_1(\theta) = \log \left( \frac{\theta}{1-\theta} \right) \) (\textbf{NB} the basis of \textbf{logistic regression}), and \( h_1(y_i) = y_i \), and (b) that \( \sum_{i=1}^{n} h_1(y_i) = s \) is sufficient for \( \theta \).
Then (102) says that the conjugate prior for the Bernoulli/binomial likelihood is

\[
p(\theta) = c (1 - \theta)^{n \tau_0} \exp \left[ \tau_1 \log \left( \frac{\theta}{1 - \theta} \right) \right]
\]

\[
= c \theta^{\alpha - 1} (1 - \theta)^{\beta - 1} = \text{Beta}(\alpha, \beta) \quad (105)
\]

for some \( \alpha \) and \( \beta \), as we’ve already seen is true.

As an example of a non-regular exponential family, suppose (as in the case study in homework 3 problem 2) that a reasonable model for the data is to take the observed values \((y_i|\theta)\) to be conditionally IID from the uniform distribution \(U(0, \theta)\) on the interval \((0, \theta)\) for unknown \( \theta \):

\[
p(y_1|\theta) = \left\{ \begin{array}{ll}
\frac{1}{\theta} & \text{for } 0 < y_1 < \theta \\
0 & \text{otherwise}
\end{array} \right\} = \frac{1}{\theta} I(0, \theta), \quad (106)
\]

where \( I(A) = 1 \) if \( A \) is true and 0 otherwise.

\( \theta \) in this model is called a range-restriction parameter; such parameters are fundamentally different from location and scale parameters (like the mean \( \mu \) and variance \( \sigma^2 \) in the \( N(\mu, \sigma^2) \) model, respectively) or shape parameters (like the degrees of freedom \( \nu \) in the \( t_\nu \) model).
The Exponential Family (continued)

(106) is an example of (99) with $c = 1, f_1(y) = 1, g_1(\theta) = \frac{1}{\theta}, h_1(y) = 0,$ and $\phi_1(\theta) =$ anything you want (e.g., 1), but only when the set $\mathcal{Y} = (0, \theta)$ is taken to depend on $\theta$.

(Truncated distributions with unknown truncation point also lead to non-regular exponential families.)

As you’ll see in homework 3, inference in non-regular exponential families is similar in some respects to the story when the exponential family is regular, but there are some important differences too.

For an example with $p > 1$, take $\theta = (\mu, \sigma^2)$ with the Gaussian likelihood:

$$l(\theta|y) = \prod_{i=1}^{n} \frac{1}{\sigma \sqrt{2\pi}} \exp \left[ -\frac{1}{2\sigma^2} (y_i - \mu)^2 \right]$$

(107)

$$= \sigma^{-n} (2\pi)^{-\frac{n}{2}} \exp \left[ -\frac{1}{2\sigma^2} \left( \sum_{i=1}^{n} y_i^2 - 2\mu \sum_{i=1}^{n} y_i + n\mu^2 \right) \right].$$
This is of the form (101) with $k = 2$, $c = (2\pi)^{-\frac{n}{2}}$, $f(y) = 1$, $g(\theta) = \sigma^{-n} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right)$, $\phi_1(\theta) = -\frac{1}{2\sigma^2}$, $\phi_2(\theta) = \frac{\mu}{\sigma^2}$, $h_1(y_i) = y_i^2$, and $h_2(y_i) = y_i$, which shows that $[h_1(y) = \sum_{i=1}^{n} y_i^2, h_2(y) = \sum_{i=1}^{n} y_i]$ or equivalently $(\bar{y}, s^2)$ is sufficient for $\theta$.

Some unpleasant algebra then demonstrates that an application of (102) leads to (88) as the conjugate prior for the Gaussian likelihood when both $\mu$ and $\sigma^2$ are unknown.
2.10 References


References (continued)


