Bayesian Methods: Introduction to Multi-parameter Models

Parameter: $q = (q_1, q_2)$

- Given Likelihood p(y|q) and prior p(q), the posterior $p(q_1, q_2 \mid y)$ is proportional to p(y|q) x p(q)

Marginal posterior

- Interested only in q_1?
    - Loss function depends on q_1 only
        - E.g. y ~ $N(m, s^2)$, both unknown. Then $q = (m, s^2)$ and L$(q, d) = (m - d(x))^2$.
    - Can show that the expected posterior loss involves the marginal posterior p$(q_1 \mid y) = \int p(q_1, q_2 \mid y) dq_2 = \int p(q_1 \mid q_2, y)\, p(q_2 \mid y) dq_2$
        - Note that the marginal posterior of q_1 is a mixture of conditional posteriors given q_2.
        - When q_2 takes discrete values (1,2,…,M), possibly denoting different models, then the posterior of q_1 is a weighted average of posteriors given each model.
            - The weights depend on the combined evidence from the prior and the data, $p(q_2 \mid y)$.
            - This is the key idea underlying Bayesian Model Averaging (BMA). Given a strong belief in the prior specification, Bayesian need not select a model. BMA would minimize expected risk.

- We can always draw samples from the joint posterior p$(q_1, q_2 \mid y)$.
    - If it is easier to sample from p($q_2 \mid y$), draw samples from this distribution, and then for each of these samples, draw from $p(q_1 \mid y, q_2)$. To obtain samples from p($q_1 \mid y$), ignore the q_2 coordinate.

Analysis for Normal Data, $N(m,s^2)$ with Non-informative prior

- Prior $p(m \mid s^2) \propto c;\ p(s^2) \propto (s^2)^{-1}$, i.e., $(m, \ln s) \sim$ Uniform

- Normal likelihood: Given (m,s), n iid observations lead to the sufficient statistics $\bar{y}, s^2 = \frac{1}{(n-1)}\sum_1^n (y_i - \bar{y})^2$ and the likelihood function

  $$(s^2)^{-n/2} \exp\{-\frac{1}{2s^2}[(n-1)s^2 + n(\bar{y}-m)^2]\}$$

- The posterior of (m,s^2), proportional to likelihood times the prior, factorizes into two parts:

  $$\exp\{-\frac{n}{2s^2}(\bar{y}-m)^2\};$$

  $$(s^2)^{-(n+2)/2} \exp\{-\frac{1}{2s^2}(n-1)s^2\}$$

- The first term represents the kernel of a Normal density with mean $\bar{y}$, and var s^2/n, except for the constant $1/\sqrt{2ps^2/n}$.
  - When using the precision notation $\{t = (1/s^2)\}$, we say that given y and t, the conditional posterior of m is Normal with mean $\bar{y}$, and precision nt.
    - Note that for a sample of size n, given (m,s), $\bar{y}$ has a Normal distribution with precision nt.

- For the marginal posterior of t or s^2:
  - We must integrate out the first term corresponding to the conditional posterior of m, given t, in the above joint posterior expression. This yields a term proportional to $\sqrt{s^2}$
    - Alternatively, the first term is proportional to $s * N(\bar{y}, s^2/n)$
    - Now, the marginal posterior density of s^2 is proportional to $(s^2)^{\frac{-(n+2)+1}{2}} \exp\{-\frac{n_1}{2s^2}\}$, where $n_1 = (n-1)s^2$.
    - Note that this corresponds to an inverse-gamma density, or a scaled inverse chi-squared. For its summary statistic, mean, median and mode, see Table A.1 in the text book (pp 574).

  - However, in order to obtain the density of t , we must account for a change of variables from s^2 to t . Since, the absolute value of ds^2/dt yields the term t^{-2}, the marginal posterior density of t is proportional to $t^{\frac{(n+1)}{2}-2}\exp\{-\frac{1}{2}n_1t\}$, which corresponds to a scaled chi-squared with (n-1) degrees of freedom and scale parameter n_1, or a $\text{Gamma}(\frac{\text{n-1}}{2},\frac{n_1}{2})$.
    - Note that, the posterior density of n_1t is a chi-squared random variable with (n-1) degrees of freedom, which is same as the sampling distribution of $n_1 = (n-1)s^2$.
    - Given this non-informative prior on the parameters, the distribution of the pivotal quantity $\frac{(n-1)s^2}{s^2}$ remained unchanged.

- The posterior distribution of (m, s^2) belongs to Normal-Inverted Gamma family, and that of (m, t) belongs to Normal-Gamma family.

- We can easily draw samples from this joint posterior by first drawing samples from a Gamma (scaled chi-squared), and then given each t , draw a sample from the conditional Normal distribution of m .

- For the Marginal Posterior of m

  - Since, the conditional posterior distribution of m given t is Normal with mean $\bar{y}$, and var s^2/n , therefore, given t
$$Z = (m-\bar{y})\sqrt{nt} \sim N(0,1) .$$
    - Since the distribution of Z doesn't depend on the conditioning variable t , (Z, t) are independent random variables.
    - Thus n_1t is chi-squared random variable with (n-1) d.f., independent of Z.
    - Hence, $\frac{Z}{\sqrt{n_1t\,/(n-1)}} = \frac{\sqrt{n}(m-\bar{y})}{s}$ ~ Student's t with (n-1) degrees of freedom. Thus, the marginal posterior of m is a t-distribution with location $\bar{y}$ and scale $s/\sqrt{n}$.

        - Note that with this non-informative prior, the sampling distribution of the pivotal quantity, t = $\frac{\sqrt{n}(\bar{y}-m)}{s}$ also has the Student's t distribution with (n-1) d.f.
        - Note that the t-distribution represents a scale-mixture of Normal random variables, when the scale has an inverted Gamma distribution.

- Posterior-predictive density of Future Observation(s)

    - In order to predict a future observable $\tilde{y}$, whose density depends on (m, s^2), we need to find the predictive density, p($\tilde{y}$ |y), when the uncertainty about the parameters (m, s^2) is given by its posterior
    - Of course, given the samples from the posterior of (m, s^2), and the density $\tilde{y} | (m, s^2)$, one can now draw samples from the joint density of $(\tilde{y}, m, t)$. Ignoring the second and third columns provides the samples from the posterior-predictive density of $\tilde{y}$.
    - However, if the future observation is also from Normal(m, s^2) population, one can easily get the analytic expression of the posterior predictive density.
        - Given (m, s^2), $\tilde{y} = m + s\,Z$, where Z is a standard Normal random variable.
        - Furthermore, $m\,|\,(y, s^2)$ is Normal with mean $\bar{y}$, and var s^2/n, it follows that $\tilde{y}\,|\,\bar{y}, s^2$ is Normal with mean $\bar{y}$, and var $s^2(1+1/n)$.
        - Hence, given $(\bar{y}, t)$, $U = \frac{t^{1/2}(\tilde{y}-\bar{y})}{\sqrt{(n+1)/n}}$ is a Normal (0,1) random variable. Furthermore, since $n_1 t$ is an independent Chi-squared random variable with (n-1) degrees of freedom, it follows that $\frac{\tilde{y}-\bar{y}}{s\sqrt{(n+1)/n}}$ has a t-distribution.
        - In other words, the posterior-predictive density of $\tilde{y}$ is a t-distribution with location $\bar{y}$ and scale $s\sqrt{(1+n)/n}$.
            - Note that if we want to predict m future observations from this same population, knowing that the $\bar{\tilde{y}}_m, s_m^2$ is the sufficient statistic, we can achieve this task by first predicting one observation from N$(m, s^2/m)$, as above, as well as one from

the predictive density of s_m^2, which can be found similarly. Now, given $\bar{\bar{y}}_m, s_m^2$, the conditional distribution of $Y_1,\cdots,Y_m$ doesn't depend on the parameters. Thus we can now draw Y's from this distribution.

- The Example on the *speed of light* is worth reading, since in this case the outliers do not satisfy the normal model, and the posterior based on this data model does not look good. In fact, in this problem, the signal to noise ratio is very small, so the model has to be really good.
  - In fact, the values of the physical constants are reviewed every five years by the '**Committee** on Data for Science and Technology' (CODATA), see, e.g., http://physics.nist.gov/cuu/Reference/contents.html and an interesting article on 'implications of non-constant velocity of light' at http://www.ldolphin.org/cdkconseq.html . CODATA evaluates the collection of observations made in the intervening five years for outliers etc, and then updates the values of physical constants. Of course, the change in a few least significant digits.

Analysis of Normal data $\underline{N(m,s^2)}$ with conjugate Normal-inverted Gamma prior

- Given the likelihood of n iid observations from Normal, the conjugate prior should also have two terms of the same form.
- It suggests the conjugate Prior: $m \mid s^2 \sim N(m_0, s^2/k_0)$. [This prior is equivalent to the posterior from a state starting with uniform prior, and drawing k_0 with observed mean m_0 when the variance is known.] In addition, $\frac{1}{s^2} = t$ ~scaled $c_{n_0}^2$ with scale $n_0 t_0 (= n_0 / s_0^2)$, or a Gamma$(\frac{n_0}{2}, \frac{n_0 t_0}{2})$.
- Note that in the conjugate prior, the two parameters are dependent, but we are assigning independent distributions to $(m/s$ and $1/s^2)$. The signal to noise ratio (m/s) is a very popular parameter in Engineering applications.
  - In effect, the prior is same as a random effect model for m, which may not be suitable in some applications. [See the textbook on this issue.]

- On multiplying the likelihood by the prior, it is easy to see that the posterior is also a Normal-Inverted Gamma form with updated parameters

$$m_n = w_n m_0 + (1-w_n)\bar{y}, \text{ where } w_n = \frac{k_0}{k_n},$$

$$k_n = k_0 + n,$$

$$n_n = n_0 + n, \text{ and}$$

$$n_n s_n^2 = n_0 s_0^2 + (n-1)s^2 + n w_n (\bar{y} - m_0)^2$$

- Again, sampling from this distribution is self-explanatory.
- Now, for the marginal posterior distribution of m, following the discussion in the non-informative prior case, it is easy to see that we get a t-distribution with location m_n and scale (s_n^2 / k_n).
- Similarly, the predictive density of a future observation can be obtained.

Analysis of Normal data with semi-conjugate prior

- In some applications, the prior on (m, s^2) may be required to be independent. In this case, the joint posterior will not factorize any more, but one can still obtain the conditional and marginal posteriors.