1 Feb 2013

eBay/Google short course: Problem set 3

1. (Bayesian transformation of variables) Continuing problem 2 from Problem Set 2, let's again consider the $n = 14$ failure time values y_i given in the statement of that problem, for which we saw that a reasonable (initial) model is based on the exponential distribution for the y_i,

$$\left\{ \begin{array}{ccc} \lambda & \sim & \Gamma^{-1}(\alpha, \beta) \\ (y_i|\lambda) & \stackrel{\text{IID}}{\sim} & \mathcal{E}(\lambda) \end{array} \right\} \Longrightarrow (\lambda|y) \sim \Gamma^{-1}(\alpha + n, \beta + n\bar{y}). \tag{1}$$

Here, as before, (i) $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, (ii) the sampling distribution for the y_i is given by

$$(y_i|\lambda) \stackrel{\text{IID}}{\sim} p(y_i|\lambda) = \left\{ \begin{array}{cc} \frac{1}{\lambda}\exp(-\frac{y_i}{\lambda}) & y_i > 0 \\ 0 & \text{otherwise} \end{array} \right\} \tag{2}$$

for some $\lambda > 0$, and (iii) the conjugate prior for λ is

$$\lambda \sim \Gamma^{-1}(\alpha, \beta) \quad \Longleftrightarrow \quad p(\lambda) = \left\{ \begin{array}{cc} \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{-(\alpha+1)}\exp\left(-\frac{\beta}{\lambda}\right) & \lambda > 0 \\ 0 & \text{otherwise} \end{array} \right\}. \tag{3}$$

In that problem I mentioned that the exponential model can either be parameterized in terms of λ or $\frac{1}{\lambda}$. In this problem we'll explore what happens when you're more interested in $\eta = g(\lambda) = \frac{1}{\lambda}$ than in λ itself.

(a) Use the **change-of-variables formula** derived below to show that the prior and posterior distributions for η are $\Gamma(\alpha, \beta)$ and $\Gamma(\alpha + n, \beta + n\bar{y})$, respectively (which justifies the name *inverse gamma* for the distribution of λ).

(b) Write out the likelihood function in terms of η instead of λ (just substitute η everywhere you see $\frac{1}{\lambda}$), and use `Maple` (or some other environment of your choosing) to plot the prior, likelihood, and posterior distributions for η on the same graph, using the data and prior values given in Problem Set 2.

(c) Use the fact that the $\Gamma(\alpha, \beta)$ distribution has mean $\frac{\alpha}{\beta}$ and variance $\frac{\alpha}{\beta^2}$ to numerically compute the prior, likelihood, and posterior means and SDs for η (you don't have to give the likelihood-maximizing summaries if you don't want to; it's enough to give results based on the likelihood-integrating approach). Is the posterior mean a weighted average of the prior and data means in this model, and if so what interpretation would you give to α and β in the $\Gamma(\alpha, \beta)$ prior for η? Explain briefly.

The change-of-variables formula. Consider a real-valued continuous random variable Y with CDF $F_Y(y) = P(Y \le y)$ and density $f_Y(y)$, related as usual to the CDF by $F_Y(y) = \int_{-\infty}^{y} f_Y(t)\,dt$ and $f_Y(y) = \frac{d}{dy}F_Y(y)$. Suppose you're interested mainly in a random variable X which is a transformed version of Y: $X = h(Y)$ for some invertible (strictly monotonic) function h. Such functions have to be either strictly increasing or decreasing; as a first case assume the former. Then the CDF of X, $F_X(x) = P(X \le x)$, satisfies

$$F_X(x) = P(X \le x) = P[h(Y) \le x] = P[Y \le h^{-1}(x)] = F_Y\left[h^{-1}(x)\right], \tag{4}$$

from which the density of X is

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{d}{dx}F_Y\left[h^{-1}(x)\right] = f_Y\left[h^{-1}(x)\right]\frac{d}{dx}h^{-1}(x) = f_Y\left[h^{-1}(x)\right]\left|\frac{d}{dx}h^{-1}(x)\right|, \tag{5}$$

the last equality holding because h, and therefore h^{-1}, are strictly increasing (and therefore both have positive derivatives). Similarly, if h is strictly decreasing,

$$F_X(x) = P(X \le x) = P[h(Y) \le x] = P[Y \ge h^{-1}(x)] = 1 - F_Y\left[h^{-1}(x)\right], \tag{6}$$

from which the density of X is

$$f_X(x) = \frac{d}{dx}F_X(x) = \frac{d}{dx}F_Y\left[h^{-1}(x)\right] = f_Y\left[h^{-1}(x)\right]\left[-\frac{d}{dx}h^{-1}(x)\right]. \tag{7}$$

But since h is strictly decreasing, so is h^{-1}, and both therefore have negative derivatives, so that

$$-\frac{d}{dx}h^{-1}(x) = \left|\frac{d}{dx}h^{-1}(x)\right|. \tag{8}$$

Thus the conclusion is that in either case

$$f_X(x) = f_Y\left[h^{-1}(x)\right]\left|\frac{d}{dx}h^{-1}(x)\right|, \tag{9}$$

which is the **change-of-variables** formula. (Since $y = h^{-1}(x)$, a simple mnemonic for this formula, using a slightly old-fashioned notation for derivatives, is $f_X(x)\,|dx| = f_Y(y)\,|dy|$.)

2. (Inference with the uniform distribution) Paleobotanists estimate the moment in the remote past when a given species became extinct by taking cylindrical, vertical core samples well below the earth's surface and looking for the last occurrence of the species in the fossil record, measured in meters above the point P at which the species was known to have first emerged. Letting $\{y_i, i = 1, \ldots, n\}$ denote a sample of such distances above P at a random set of locations, the model $\boxed{(y_i|\theta) \overset{\text{IID}}{\sim} \text{Uniform}(0, \theta)\ (*)}$ emerges from simple and plausible assumptions. In this model the unknown $\theta > 0$ can be used, through carbon dating, to estimate the species extinction time. This problem is about Bayesian inference for θ in model $(*)$, and it will be seen that some of our usual intuitions (derived from the Bernoulli, Poisson, and Gaussian case studies) do not quite hold in this case.

The marginal sampling distribution of a single observation y_i in this model may be written

$$p(y_i|\theta) = \left\{ \begin{array}{cc} \frac{1}{\theta} & \text{if } 0 \le y_i \le \theta \\ 0 & \text{otherwise} \end{array} \right\} = \frac{1}{\theta}\, I\left(0 \le y_i \le \theta\right), \tag{10}$$

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

(a) Use the fact that $\{0 \le y_i \le \theta \text{ for all } i = 1, \ldots, n\}$ if and only if $\{m = \max(y_1, \ldots y_n) \le \theta\}$ to show that the likelihood function in this model is

$$l(\theta|y) = \theta^{-n} I(\theta \ge m). \tag{11}$$

Briefly explain why this demonstrates that m is sufficient for θ in this model.

(b) As we saw in part 2 of the lecture notes (pages 17–18), the maximum likelihood estimator (MLE) of a parameter θ is the value of θ (which will be a function of the data) that maximizes the likelihood function, and this maximization is usually performed by setting the derivative of the likelihood (or log likelihood) function to 0 and solving. Show by means of a rough sketch of the likelihood function in (a) that m is the maximum likelihood estimator (MLE) of θ, and briefly explain why the usual method for finding the MLE fails in this case.

(c) A positive quantity θ follows the **Pareto** distribution (written $\theta \sim \text{Pareto}(\alpha, \beta)$) if, for parameters $\alpha, \beta > 0$, it has density

$$p(\theta) = \left\{ \begin{array}{cc} \alpha\, \beta^{\alpha}\, \theta^{-(\alpha+1)} & \text{if } \theta \geq \beta \\ 0 & \text{otherwise} \end{array} \right\}. \tag{12}$$

This distribution has mean $\frac{\alpha\beta}{\alpha-1}$ (if $\alpha > 1$) and variance $\frac{\alpha\beta^2}{(\alpha-1)^2(\alpha-2)}$ (if $\alpha > 2$).

With the likelihood function viewed as (a constant multiple of) a density for θ, show that equation (11) corresponds to the Pareto$(n-1, m)$ distribution. Show further that if the prior distribution for θ is taken to be (12), under the model $(*)$ above the posterior distribution is $p(\theta|y) = \text{Pareto}\,[\alpha + n, \max(\beta, m)]$, thereby demonstrating that the Pareto distribution is conjugate to the Uniform$(0, \theta)$ likelihood.

(d) In an experiment conducted in the Antarctic in the 1980s to study a particular species of fossil ammonite, the following was a linearly rescaled version of the data obtained, in ascending order: $y = (y_1, \ldots, y_n) = (0.4, 1.0, 1.5, 1.7, 2.0, 2.1, 2.8, 3.2, 3.7, 4.3, 5.1)$. Prior information equivalent to a Pareto prior specified by the choice $(\alpha, \beta) = (2.5, 4)$ was available. Plot the prior, likelihood, and posterior distributions arising from this data set on the same graph, explicitly identifying the three curves, and briefly discuss what this picture implies about the updating of information from prior to posterior in this case.

(e) Make a table summarizing the mean and standard deviation (SD) for the prior (Pareto(α, β)), likelihood (Pareto$(n-1, m)$), and posterior (Pareto$[\alpha + n, \max(\beta, m)]$) distributions, using the (α, β) choices and the data in part (d) above (as in problem 1, it's enough to do this using the likelihood-integrating approach). In Bayesian updating the posterior mean is usually (at least approximately) a weighted average of the prior and likelihood means (with weights between 0 and 1), and the posterior SD is typically smaller than either the prior or likelihood SDs. Are each of these behaviors true in this case? Explain briefly.

(f) You've shown in (c) that the posterior for θ based on a sample of size n in model $(*)$ is $p(\theta|y) = \text{Pareto}\,[\alpha + n, \max(\beta, m)]$. Write down a symbolic expression for the posterior variance of θ in terms of (α, β, m, n). When considered as a function of n, what's unusual about this expression in relation to the findings in our previous case studies in this course? Explain briefly.

3. (Inference for the variance in the Gaussian model with known mean) As we saw in problem 4 of Problem Set 1, American football experts provide a *point spread* for every football game before it occurs, as a measure of the difference in ability between the two teams (and taking account of where the game will be played). For example, if Denver is

a 3.5–point favorite to defeat San Francisco, the implication is that betting on whether Denver's final score minus 3.5 points exceeds or falls short of San Francisco's final score is an even-money proposition. Figure 1 below (based on data from Gelman et al. 2003) presents a histogram of the differences d = (actual outcome – point spread) for a sample of $n = 672$ professional football games in the early 1980s, with a normal density superimposed having the same mean $\bar{d} = 0.07$ and standard deviation (SD) $s = 13.86$ as the sample. You can see from this figure that the model $(D_i|\sigma^2) \overset{\text{IID}}{\sim} N(0, \sigma^2)$ is reasonable for the observed differences d_i (at least as a starting point in the modeling).

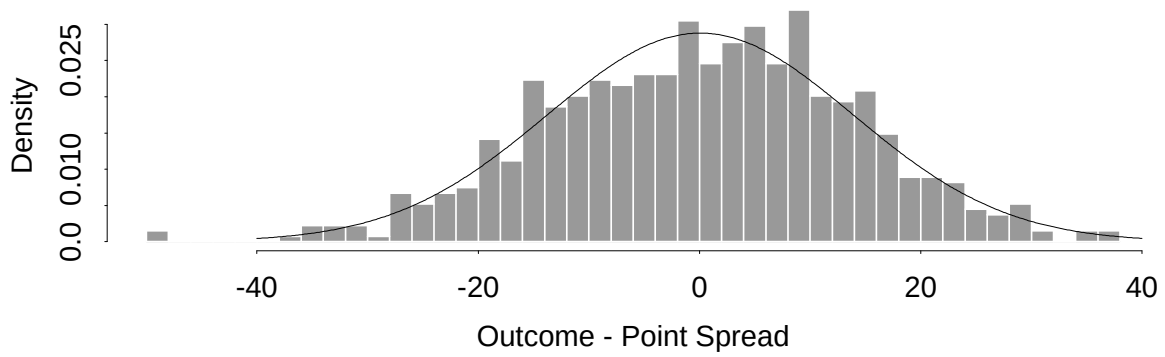

Figure 1. *Differences d_i between observed and predicted American football scores, 1981–1984.*

(a) Write down the likelihood and log likelihood functions for σ^2 in this model. Show that $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} d_i^2$, which takes the value 191.8 with the data in Figure 1, is both sufficient and the maximum likelihood estimator (MLE) for σ^2. Plot the log likelihood function for σ^2 in the range from 160 to 240 with these data, briefly explaining why it should be slightly skewed to the right.

(b) The conjugate prior for σ^2 in this model is the *scaled inverse chi-square* distribution,

$$\sigma^2 \sim \chi^{-2}(\nu_0, \sigma_0^2), \quad \text{i.e.,} \quad p(\sigma^2) = c \left(\sigma^2\right)^{-\left(\frac{\nu_0}{2}+1\right)} \exp\left(-\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right), \tag{13}$$

where ν_0 is the prior sample size and σ_0^2 is a prior estimate of σ^2. In an attempt to be "non-informative" people sometimes work with a version of (13) obtained by letting $\nu_0 \to 0$, namely $p(\sigma^2) = c_0 \left(\sigma^2\right)^{-1}$. The resulting prior is *improper* in that it integrates to ∞, but it turns out that posterior inferences will be sensible nonetheless (even with sample sizes as small as $n = 1$). Show that with this prior, the posterior distribution is $\chi^{-2}(n, \hat{\sigma}^2)$.

Figure 2 below plots the prior, likelihood, and posterior densities on the same graph using the data in Figure 1 and taking $c_0 = 2.5$ for convenience in the plot. Get `R` (or some equivalent environment) to reproduce this figure (**NB** `Maple` has a hard time doing this). You'll need to be careful to use the correct normalizing constant c in (13), which can be found either in the lecture notes or in Appendix A of Gelman et al. (2003); and because the data values in this example lead to astoundingly large and small numbers on the original scale, it's necessary to do all possible computations on the log scale and wait to transform back to the original scale until the last possible moment (you'll need to use the built-in function `lgamma` in `R`, or something like it in your favorite environment). Explicitly identify

the three curves, and briefly discuss what this plot implies about the updating of information from prior to posterior in this case.

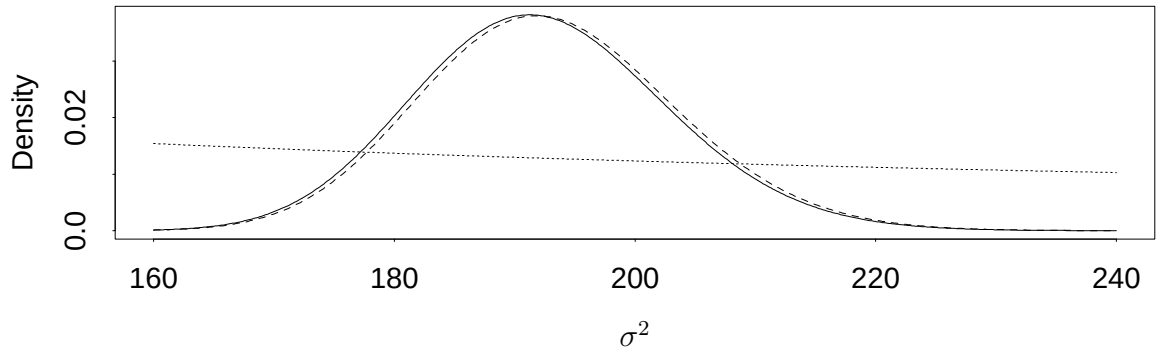

Figure 2. *Prior, likelihood, and posterior densities with the football data of Figure 1.*

4. (The gambler's ruin) Consider a gambler at a casino who at each play of a game has probability $0 < p < 1$ of winning \$1 and probability $(1 - p)$ of losing \$1. If the successive plays of the game are assumed independent, the question this problem addresses is as follows: what is the probability P that if she (the gambler) starts with $\$M > \$0$ she will *break the bank* (reach $\$N > \$M$, for integer M and N; here $\$N$ represents the initial capital of the casino against which she's playing[1]) before *going broke* (reaching \$0)?

(a) If we let Y_t denote her fortune after the tth play of the game, explain why the process $\{Y_t\}$ is a Markov chain on the state space $\{\$0, \$1, \ldots, \$N\}$, and identify the possible states the process could be in at times $t = 0, 1, \ldots$.

(b) My intent is that this problem should be a somewhat playful environment within which you can learn more about Markov chains than you already know (and the grading of the problem will be accordingly liberal). Therefore, using whatever combination you like of {simulation (`R` is a good language for this), looking around on the web, reading probability books, etc.}, see how much progress you can make on the basic question posed at the beginning of the problem. A fully satisfying mathematical answer to the question would be symbolic in p, M, and N, but you'll get nearly full credit for doing a good job of answering it for a few (well-chosen) specific values of these quantities and speculating about the nature of the dependence of P on p, M, and N. Explore the sensitivity of P to small changes in p, M, and N: on which of these quantities does P depend most sensitively?

(c) Let $N \to \infty$ and show that under these conditions, if $p > \frac{1}{2}$ there is a positive probability (specify it if you can) of the gambler's fortune increasing indefinitely, but if $p \leq \frac{1}{2}$ she will go broke with probability 1 against an infinitely rich adversary (this last fact is not surprising for $p < \frac{1}{2}$, but what about $p = \frac{1}{2}$?).

[1]For the sake of this problem let's pretend that once she reaches $\$N$ the casino judges that it has lost enough money to her that it does not wish to continue playing against her, which is what "breaking the bank" means.