Econ. 552b Spring 1996 Chris Sims

Take-Home Answers

1. A prior p.d.f. that is uniform over $0 < a,b < 1$, $a + b < 1$, and integrates to a probability of .8 must have height 1.6, as the area of the region to be integrated over is .5. The height of a uniform prior p.d.f. along $a + b = 1$, $0 < a < 1$ that integrates to .2 depends on what the index of integration is. It may seem natural that since the length of the line integrated over is $\sqrt{2}$, the p.d.f. height should be $.2/\sqrt{2}$. But actually it is probably most convenient to parameterize the line by setting $b = 1-a$ (or vice versa) and then to integrate with respect to a. In that case the appropriate p.d.f. height for the prior is just .2.

The likelihood has the shape of a multivariate *t*. The problem didn't say how to treat the constant term *g*, but is was natural to assume this had a flat prior, so that one just integrates it out to get the posterior on *a* and *b*, which will still be multivariate *t*, in form. A somewhat less natural way to handle this was to pretend *g* was known and fixed. If z=(*x*,*y*) has a bivariate *t* distribution with *n*-3 degrees of freedom (as would emerge as the likelihood shape from our regression model with *n* observations and three estimated parameters) with covariance matrix Σ, its p.d.f. is proportional to

$$(n-3+z\Sigma^{-1}z)^{-(n-1)/2} \tag{A.1}$$

For later reference, the formula for the p.d.f. of a *d*-dimensional vector *t* with mean zero, covariance matrix Σ, and a multivariate *t* distribution with degrees of freedom h, is (adapted from Robert, p.382)

$$\frac{\Gamma\left(\frac{h+d}{2}\right)}{\Gamma\left(\frac{h}{2}\right)} h^{h/2}|\Sigma|^{-\frac{1}{2}} p^{-d/2}\left(h+t'\Sigma^{-1}t\right)^{-(h+d)/2} . \tag{A.2}$$

I was surprised to note that Gelman, *et al*, does not contain the full formula with scale factor, only the last part, which is all that depends on *t*. For this problem you do need the full formula, and allowances were made for the fact that it is hard to derive and not in one of the two standard textbooks for the course. (It is in other widely available references, though.) If we set $\hat{m} = 1-\hat{a}-\hat{b}$, and let $x = a-\hat{a}$, $y = b-\hat{b}$, then along the line $a+b=1$, $y = \hat{m}-x$, and as a function of*x*, (A.1) can be rewritten as

$$\left(n-3+x^2 s^{11}+2(\hat{m}-x)x s^{12}+(\hat{m}-x)^2 s^{22}\right)^{-(n-1)/2} . \tag{A.3}$$

While messy looking, (A.3) is nonetheless a quadratic function of *x* raised to an integral multiple of $-\frac{1}{2}$, so it is in the form of a univariate *t*. We need only (!) find what its scale factor is to integrate it. Collecting terms and completing the square, (A.3) becomes

$$\left(n-3+\hat{m}^2 s^{22}+x^2\left(s^{11}-2s^{12}+s^{22}\right)-2\hat{m}\cdot\left(s^{22}-s^{12}\right)x\right)^{-(n-1)/2}$$

$$=\left(n-3+\hat{m}^2\frac{s^{11}s^{22}-\left(s^{12}\right)^2}{s^{11}-2s^{12}+s^{22}}+\left(x-\hat{m}\frac{s^{22}-s^{12}}{s^{11}-2s^{12}+s^{22}}\right)^2\left(s^{11}-2s^{12}+s^{22}\right)\right)^{-(n-1)/2} \quad \text{(A.4)}$$

This is proportional to the p.d.f. of a univariate t with n - 2 degrees of freedom and mean and variance parameters

$$m=\hat{m}\frac{s^{22}-s^{12}}{s^{11}-2s^{12}+s^{22}} \quad \text{(A.5)}$$

$$s^2=\frac{n-3+\hat{m}^2\dfrac{s^{11}s^{22}-\left(s^{12}\right)^2}{s^{11}-2s^{12}+s^{22}}}{(n-2)\left(s^{11}-2s^{12}+s^{22}\right)}\ . \quad \text{(A.6)}$$

However (A.4) represents the standard form scaled by the factor

$$\left(\frac{1+\hat{m}^2\dfrac{s^{11}s^{22}-\left(s^{12}\right)^2}{s^{11}-2s^{12}+s^{22}}}{n-2}\right)^{-(n-1)/2}(n-2)^{-(n-2)/2}sp^{1/2}\frac{\Gamma\left(\dfrac{n-2}{2}\right)}{\Gamma\left(\dfrac{n-1}{2}\right)} \quad \text{(A.7)}$$

The product of (A.7) with the scale factor in (A.2) specialized to our case of n - 3 degrees of freedom, covariance matrix Σ, would provide the posterior probability of $a + b = 1$ if we had a flat prior, the height of the prior p.d.f. on this line were 1, and there were no restrictions on the range of the continuous part of the posterior distribution. To specialize to our case, we must consult a t table to find the probability that a t with n - 2 degrees of freedom lies in $(-m/s,(1-m)/s)$, with m and s defined by (A.5) and (A.6), and multiply that by .2, getting a number we'll call p_1. Then we must simulate a bivariate t with n - 3 degrees of freedom and covariance matrix Σ, finding the proportion of draws that fall in the region $t_1 > -\hat{a}$, $t_2 > -\hat{b}$, $t_1 + t_2 < \hat{m}$, and multiply this probability by 1.6, getting a number we'll call p_2. The posterior probability of $a + b = 1$ is then $p_1/(p_1 + p_2)$.

An alternative approach could have been based on Gibbs sampling. For any given value of a, the conditional likelihood as a function of b has the form of a univariate t with n - 2 degrees of freedom, though with mean and variance parameters that will differ for every a. The conditional distribution of b is then a truncated univariate t together with a discrete weight on $b = 1 - a$. The probability in the truncated t region and on the discrete point can be computed analytically. A monte carlo draw from this distribution is then easy to form. Conditioning on this draw for b, we then do the same sort of thing for a, etc. The posterior probability on $a + b = 1$ is then estimated as the proportion of draws for which the equality holds exactly.

The algebra here was forbidding in spots, but it was disappointing that most answers did not even get to the point of correctly describing what needed to be calculated.

2. This was an essentially standard stopping rule problem, which few answers seemed to recognize. Since sufficient statistics were being reported, it was possible to construct the likelihood, and the principle in stopping problems is that if the rule for stopping makes sample size a function of the data sequence, the likelihood function has the same form as if the sample size is non-random. Thus in the first example Bayesian conclusions are the same as if the sample size had been deterministic, whether observed sample size is 87 or 150. The second example is trickier. The stopping rule is a function of the data only. We are being given only the results associated with the larger of two estimates. Certainly we would not treat the OLS estimate and its standard error as providing the same kind of evidence as if there were a single sample being analyzed. The interesting question is whether our inference here would be different from the case where, with a deterministic sample size, just the larger of two OLS estimates were reported to us. It might seem that this should be the case, but since we are not being given sufficient statistics, the usual argument that the likelihood function has the same form for fixed *N* as for an *N* determined by the data does not apply. I'm pretty sure that it actually does not apply, from considering simple binomial examples, but haven't proved it.

No exam gave a careful discussion of any well-defined classical inferential procedure for the problem. One could, for example, consider a likelihood-ratio test for the null of $b = 0$. Because the likelihood function itself depends only on the *t*-ratio or (if you take the variance as known) the value of b, a likelihood ratio test will also pay no attention to the sample size, only to the usual test statistic. However, the classical distribution theory for this test statistic will not be standard. If the variance is unknown, the distribution for the *t* ratio depends on the unknown variance in a nasty way, making a test of the null of $b = 0$ without a restriction on the variance parameter difficult to handle. In the simple special case of $X \equiv 1$, so we are estimating a mean, and with $s = 1$, the distribution of the sample mean at the stopping point is a mixture of distributions concentrated on $\hat{b} > 1$, contributed by the possibility of an early stopping time, with a distribution that is skewed toward negative $\hat{b}$'s, which is the form of the distribution conditional on stopping at $T = 150$. The negative skewness comes from the fact that on sample paths that lead to a $\hat{b}$ close to 1 at $T = 150$ if we ignored the stopping rule, the odds are good that the data collection would have stopped before $T = 150$. Because we are mixing these two distributions with opposite skewness, the question of whether a classical hypothesis test is more or less likely to reject the null with a given observed $\hat{b}$ than would be a classical test that ignored the stopping rule is quite subtle. No answer recognized that the fact that the sampling *could* go on to $T = 150$, and that at that point test statistics would be "biased" *downward* under a null of $b = 0$, has strong effects on classical inference for situations where actual T is below 150.

3. Since

$$vec(B) = -(I \otimes \Gamma)vec(\Pi) \ , \tag{A.8}$$

the Jacobian of the transformation from Γ,B to Γ,Π is $|\Gamma|^q$, where q is the number of columns in Π and B, i.e. the number of predetermined variables. This means that the joint p.d.f. for Γ,Π is $f(\Gamma,-\Gamma\Pi)\cdot|\Gamma|^q d\Gamma d\Pi$. Contrary to what the problem statement implies, this transformation does not introduce any non-smoothness into the density. I had the signs of my exponents mixed up when I formulated the problem. What it does instead is introduce zeros in the p.d.f. at singularities in Γ. In fact, it is in the case where we start with a smooth prior $g(\Gamma,\Pi)$ on Γ,Π and derive the implied prior on Γ,B that the Jacobian becomes $|\Gamma|^{-q}$ and we end up with singularities in the prior p.d.f. on Γ,B. Because $\Pi=\Gamma^{-1}B$ explodes as Γ approaches singularity for any given non-singular B, we don't generally have singularity for $g(\Gamma,-\Gamma^{-1}B)|\Gamma|^{-q}$ at every point where Γ is singular, but we do have singularity at any point where B has a rank deficiency matching that in Γ, so that $\Gamma^{-1}B$ is bounded despite the singularity in Γ. This probably does not make sense. If Γ approaches singularity, we do not expect Π to remain nicely behaved, because of its origin as $-\Gamma^{-1}B$. A smooth prior on Γ and Π jointly implies unreasonably that we think it likely that Γ and B approach row-rank deficiency together in this way. In a supply and demand model, for example, a backward-bending supply curve of nearly the same slope as the demand curve implies that small shifts in supply or demand are likely to produce large changes in quantity and price. To make the prior smooth in Γ and Π implies that instead the coefficients on the shifting variables are likely to be close in these circumstances, so that anything that shifts demand is likely to shift supply by almost the same amount, resulting in little change in price and quantity.

On the other hand, the $|\Gamma|^q$ term in the implied prior on Γ,Π tends to give this p.d.f. fat tails -- very large Π's associated with near-singular Γ's are fairly likely. When Π is large, the implied amount of variability in y is large. This means that we are putting fairly high probability on observing erratic behavior in the y's, and this may not be reasonable.

Thus there is no single right answer to this question; a good answer would have discussed some of the considerations above.

4. The likelihood is

$$L(Y;b,s)=s^{-T}\exp\left(\sum_{t=1}^{T}\left(\frac{-\left(\log(y_t-X_tb+1)+\frac{s^2}{2}\right)^2}{2s^2}-\log(y_t-X_tb+1)\right)\right). \qquad (9)$$

This is derived by substituting into the normal density and using the Jacobian of the log transformation. Because of time pressure (I'm about to leave town for two weeks) I can't give you a complete argument here. However the crucial steps are as follows. First we need consistency. This is a hard argument, and it would have been OK just to assume consistency. OLS is easily shown to be consistent, which implies that any Bayesian posterior mean from a proper prior is consistent, but the MLE is here not a posterior mean under a proper prior. For

asymptotic normality, we must verify that the high-order terms in a Taylor expansion of the log likelihood around the MLE are of vanishing importance for the likelihood level in large samples. The linear term in the log likelihood is not a problem, which some of you did not realize. The MLE by construction has no linear term in its Taylor expansion. The problem is only appropriately to bound the high order terms. This could be done by examining the behavior of third derivatives, for example, or else by directly examining the dependence of the second derivative on the deviation from the MLE. If you assume the X's are bounded, the third derivatives are bounded except in the neighborhood of $e = -1$. Since the p.d.f. goes rapidly to zero in this range, it is not hard to construct an argument that the third derivatives are bounded with high probability, and this in turn justifies the Taylor approximation for parameter values not too far from the MLE. A complete argument would then also show that for parameter values more than any fixed distance, say d, from the MLE, the likelihood becomes small relative to its value near the MLE.