

ECON 623
Answers to Problem Set 2
John Rust, Fall 2010

I. Read chapters 3, 5, 8, 9 and 10 of Morris Degroot, *Optimal Statistical Decisions* McGraw Hill, 1970.

Answer Everyone is assumed to have done this and we give points for this as a free bonus.

II. Consider *Bayesian estimation* of the parameter θ where $\{X_1, \dots, X_N\}$ are N *i.i.d.* observations from a density $f(x|\theta)$ where θ is a $K \times 1$ vector of unknown parameters to be estimated. Here we assume that you have a *prior distribution* over θ represented by a density $p(\theta)$.

1. Write a formula for the *posterior density* of θ given the observations $\{X_1, \dots, X_N\}$.

Answer Using *Bayes Rule* if the prior density is $p(\theta)$ and we use the shorthand $f(X|\theta)$ as the conditional density of the data X given θ , then the posterior is $p(\theta|X)$ given by

$$p(\theta|X) = \frac{f(X|\theta)p(\theta)}{\int_{\theta'} f(X|\theta')p(\theta')d\theta'}. \quad (1)$$

In the case of *i.i.d.* sampling we have for N observations from the parametric density $f(x|\theta)$ the relevant conditional density for $X = (X_1, \dots, X_N)$ given θ is

$$f(X|\theta) \equiv \prod_{i=1}^N f(x_i|\theta). \quad (2)$$

2. Suppose $f(x|\theta)$ is a normal distribution with an unknown mean μ but a *known* variance $\sigma^2 = 1$. Suppose your prior distribution for μ is Normally distributed with mean of 1 and a variance of 1. What is the posterior distribution for θ in this case?

Answer The Normal family of distributions is an example of a *conjugate prior family* as discussed in the DeGroot chapters. If $p(\theta)$ is a normal density, then if the data X are normally distributed, then the posterior $p(X|\theta)$ will also be a normal distribution. Specifically, in this case the only unknown parameter is $\theta = \mu$, the unknown mean of the normal distribution for the data. If $p(\theta)$ is such that the prior beliefs are $\theta \sim N(1, 1)$ (i.e. if your beliefs about the unknown mean μ can be represented as a normally distributed random variable with expected value 1 and variance 1), then the posterior distribution $p(\theta|X)$ (where $X = (X_1, \dots, X_N)$ is a normal distribution with variance $\sigma^2(X)$ given by

$$\sigma^2(X) = \frac{1}{1+N} \quad (3)$$

and mean $\mu(X)$ given by

$$\mu(X) = \sigma^2(X) \left(1 + \sum_{i=1}^N X_i \right). \quad (4)$$

Notice that if we have no data, $N = 0$, then the mean and variance of the posterior distribution reduce to $\mu(X) = 1$ and $\sigma^2(X) = 1$, i.e. if we observe no data the posterior distribution is just our prior distribution which is $N(1, 1)$. Note also that as $N \rightarrow \infty$ we have that $\sigma^2(X) \rightarrow 0$ and with

probability 1 (by the Strong Law of Large Numbers) $\mu(X) \rightarrow \mu^*$, where μ^* is the true parameter of the normal density for the data, i.e. $X_i \sim N(\mu^*, 1)$. This implies that the posterior $p(\theta|X)$ converges in distribution to a unit mass on the μ^* , i.e. the Bayesian eventually learns the true unknown mean μ^* and is absolutely certain about it as $N \rightarrow \infty$.

- Use what you know so far about asymptotic theory to show what happens to the posterior distribution as $N \rightarrow \infty$. Does the posterior distribution converge in some sense to the “true” parameter θ^* (the true parameter is the parameter that “nature” uses for the “data generating mechanism”, that is the data are assumed to be draws from the density $f(x|\theta^*)$). Do you need to make any assumptions on your prior $p(\theta)$ when you determine these asymptotic properties of the Bayesian estimator? In particular, what happens if θ^* is not in the support of $p(\theta)$: can the posterior distribution converge to the true θ^* in this case?

Answer The result in part 3, namely that in the normal case, the Bayesian posterior distribution converges to a point mass on the true unknown parameter generating the data, holds more generally. Suppose in the general case that θ^* is the true unknown parameter generating the observed data $X = (X_1, \dots, X_N)$, so that as above the density for a single observation X_i is $f(X_i|\theta^*)$ and the observations are *IID* observations from this density (which is unknown to us since we don’t know θ^* although we do have prior beliefs about its possible value given by the prior density $p(\theta)$). Let $p(\theta|X_1, \dots, X_N)$ be the posterior density, and assume that θ^* is in the *support* of the prior. In the simplest terms this means that $p(\theta^*) > 0$ (so the Bayesian believes that there is a positive density that the true parameter is θ^*), or more generally it means that we can find a little ball of radius $\varepsilon > 0$ centered on θ^* , $B(\theta^*, \varepsilon)$ and the *prior probability* that θ falls into this ball is positive:

$$\Pr\{B(\theta^*, \varepsilon)\} = \int_{\theta'} I\{\theta' \in B(\theta^*, \varepsilon)\} p(\theta') d\theta' > 0, \quad (5)$$

for some $\varepsilon > 0$. Now to show that the posterior distribution converges to a point mass at the true value, θ^* , it suffices to show that *for any* $\varepsilon > 0$ the *posterior probability* that θ falls in the ball $B(\theta^*, \varepsilon)$ *converges to 1 with probability 1*. That is, we have with probability 1,

$$\forall \varepsilon > 0 \quad \lim_{N \rightarrow \infty} \int_{\theta'} I\{\theta' \in B(\theta^*, \varepsilon)\} p(\theta'|X_1, \dots, X_N) d\theta' = 1. \quad (6)$$

Since $\varepsilon > 0$ is arbitrary, it follows that the posterior density $p(\theta|X_1, \dots, X_N)$ converges in distribution to a unit mass on the true parameter θ^* (note also that the posterior densities are *random densities* since these densities depend on the data (X_1, \dots, X_N) and the data are random variables).

To show the result above, it suffices to show that for any θ that is more than a positive distance ε away from the true θ^* we have

$$\lim_{N \rightarrow \infty} p(\theta|X_1, \dots, X_N) = 0. \quad (7)$$

(thus, if the density is zero in the limit for any θ further than ε from θ^* , the probability of being within ε of θ^* , i.e. of lying in the ball $B(\theta^*, \varepsilon)$, must be converging to 1).

To do this, we show that the *posterior ratio* $p(\theta|X_1, \dots, X_N)/p(\theta^*|X_1, \dots, X_N)$ converges to 0 with probability 1 when $\theta \neq \theta^*$ (note that if $\theta = \theta^*$ then clearly the posterior ratio equals 1). Note that the posterior ratio is closely related to the *likelihood ratio*

$$\frac{p(\theta|X_1, \dots, X_N)}{p(\theta^*|X_1, \dots, X_N)} = \left(\frac{\prod_{i=1}^N f(X_i|\theta)}{\prod_{i=1}^N f(X_i|\theta^*)} \right) \left(\frac{p(\theta)}{p(\theta^*)} \right) \quad (8)$$

By our assumption that θ^* is in the support of the prior, $p(\theta^*) > 0$, so the ratio of the prior densities is a finite positive number (possibly zero). If the ratio of the priors is strictly positive, then we can prove our result by showing that the likelihood ratio (the first term in parens in the equation above) tends to 0 with probability 1 as $N \rightarrow \infty$. This is a consequence of the *consistency of the likelihood ratio statistic* (i.e. the likelihood ratio test statistic will eventually reject any hypothesis test $H_0 : \theta = \theta_0$ for any $\theta_0 \neq \theta^*$).

However a more direct way to see that this is true is to note that we can write the likelihood ratio, $\text{LR}(\theta, \theta^*)$ as

$$\text{LR}(\theta, \theta^*) \equiv \exp \left(\log \left(\frac{\prod_{i=1}^N f(X_i|\theta)}{\prod_{i=1}^N f(X_i|\theta^*)} \right) \right) \quad (9)$$

Now note that using properties of the log function we can rewrite this as

$$\text{LR}(\theta, \theta^*) = \exp \left(N \left(\frac{1}{N} \sum_{i=1}^N \log f(X_i|\theta) - \frac{1}{N} \sum_{i=1}^N \log f(X_i|\theta^*) \right) \right). \quad (10)$$

Note that by the Strong Law of Large Numbers, for any θ we have with probability 1,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N \log f(X_i|\theta) = \int \log(f(x|\theta))f(x|\theta^*)dx \quad (11)$$

and recall by the *Information Inequality* that if θ^* is *identified* (i.e. for any $\theta \neq \theta^*$, there is positive probability that $f(\tilde{x}|\theta) \neq f(\tilde{x}|\theta^*)$), then we have

$$\int \log(f(x|\theta))f(x|\theta^*)dx < \int \log(f(x|\theta^*))f(x|\theta^*)dx. \quad (12)$$

It follows that with probability 1, we have

$$\lim_{N \rightarrow \infty} \left(\frac{1}{N} \sum_{i=1}^N \log f(X_i|\theta) - \frac{1}{N} \sum_{i=1}^N \log f(X_i|\theta^*) \right) < 0, \quad (13)$$

which implies by equation (9) that with probability 1 we have

$$\lim_{N \rightarrow \infty} \text{LR}(\theta, \theta^*) = 0. \quad (14)$$

Since the likelihood ratio converges to zero with probability 1, so does the posterior ratio, and this implies that in the limit the posterior distribution places zero probability on any value of θ outside an arbitrarily small ball of radius ε about θ^* . But since probability densities must integrate to 1, this implies that in the limit the posterior places all mass (a unit mass) on the single point θ^* , i.e. the posterior converges in distribution to a unit mass on θ^* . The interpretation of this result is that *as long as the Bayesian does not rule out the possibility that the true θ could equal the actual value θ^* , the Bayesian will ultimately learn the true value θ^* with probability 1 as he/she gets more and more data and $N \rightarrow \infty$. A simpler way to say this is that the Bayesian posterior is *consistent*.*

III. The *support* of a distribution F on a metric space Θ is the smallest closed subset of X that has probability 1 under the distribution F . Prove that if Π is a prior distribution over a parameter space that is some Euclidean space or metric space Θ , that if a statistician observes data x from a density (data generating process) $f(x|\theta^*)$, and forms a posterior distribution $\Pi(\theta|x)$, the support of the posterior distribution is a subset of the support of the prior distribution.

Answer This follows from the Bayes Rule formula (1) in problem II-1. An easy way to show this is to show that if $p(\theta|X) > 0$ then $p(\theta) > 0$, i.e. if the Bayesian puts a positive posterior density on some θ in the parameter space, then it must be the case that the Bayesian had positive prior density for this θ as well. This implies that the support of the posterior distribution is a subset of the prior distribution.

IV. Section 10.1 of DeGroot discusses the concept of an *improper prior* and gives an example of a normal distribution with unknown mean ω . The improper prior is taken to be a uniform distribution over the real line. In what sense is this an improper distribution? Even if this prior is improper, if we follow mechanically the equation for Bayes Rule and derive the posterior distribution, prove that with an improper “flat prior” for ω , the posterior distribution after observing N observations (x_1, \dots, x_N) from a distribution $N(\omega, 1)$ is a normal distribution with mean \bar{X}_N , the sample mean,

$$\bar{X}_N = \frac{1}{N} \sum_{i=1}^N x_i \quad (15)$$

If we try to make the prior a proper prior by, say, limiting the prior distribution to be uniform on a compact interval, say $[-b, a]$, is it then possible for the posterior distribution to be normally distributed with mean \bar{X}_N ?

Answer A uniform distribution cannot have support over the entire real line for the following reason. A uniform density is a density that satisfies $p(\omega) = k$ for some positive constant $k \geq 0$. If $k = 0$ then the density cannot integrate to 1, so we must have $k > 0$. However if the support of θ is the entire real line (or more generally any unbounded subset of R) then the integral of this uniform density equals ∞ which again does not equal 1, the value required for a *proper density function*. However mechanically inserting this improper uniform prior into the equation for the posterior distribution in equation (1) we see that the prior cancels from numerator and denominator. In this case the likelihood of the data is

$$\prod_{i=1}^N f(x_i|\omega) = \frac{1}{(2\pi)^{N/2}} \exp\left\{-\sum_{i=1}^N (x_i - \omega)^2/2\right\} \quad (16)$$

Using the trick of “completing the square” in the sum in the right hand side of this expression we can re-write this as

$$\prod_{i=1}^N f(x_i|\omega) = \left[\frac{1}{\sqrt{N}(2\pi)^{(N-1)/2}} \exp\left\{\sum_{i=1}^N (x_i - \bar{x}_i^2)\right\} \right] \left[\frac{\sqrt{N}}{\sqrt{2\pi}} \exp\left\{-(\omega - \bar{X}_N)^2/(2/N)\right\} \right]. \quad (17)$$

Notice that the second factor in brackets on the right hand side of equation (17) is a density of a normal distribution for ω with mean \bar{X}_N and variance $1/N$. Thus, we have rewritten the expression for $\prod_{i=1}^N f(x_i|\omega)$ (the likelihood for the (x_1, \dots, x_N) conditional on ω) as a (finite) proportional factor that is only a function of the data and not ω (the first expression in brackets on the right hand side of equation (17)), times a normal density for ω with mean \bar{X}_N and variance $\frac{1}{N}$. This implies that even if we integrate over ω using an improper prior over the entire real line, the integral will still be finite. Normally with an improper prior, we would expect that the denominator in the equation for $p(\omega|x_1, \dots, x_n)$ in equation (1) would equal infinity, but with the trick above, we see that the integral in the denominator can be written as

$$\int_{\theta'} \prod_{i=1}^N f(x_i|\omega) d\theta' = \frac{\sqrt{2\pi}}{\sqrt{N}} \frac{1}{(2\pi)^{N/2}} \exp\left\{\sum_{i=1}^N (x_i - \bar{x}_i^2)\right\}, \quad (18)$$

which is indeed finite. Indeed, the first factor in brackets on the right hand side of equation (17) above cancels out in the numerator and denominator of the equation for $p(\omega|x_1, \dots, x_n)$ in (1), so it follows that in this case, *the use of an improper prior still leads to a proper posterior density for ω* . In fact, applying the same arguments above to the numerator expression for the posterior (using the general formula (1)), it is easy to see now that the posterior density for the improper flat prior is a normal distribution with mean \bar{X}_N and variance $\frac{1}{N}$.

Another way to see this result is to treat an improper flat prior as a limit of proper Gaussian priors with a variance that tends to ∞ . As the variance of a normal distribution gets larger and larger, the density becomes flatter and flatter, and it also converges to zero. Consider the case where the prior parameters of the prior for ω are ω_0 and σ_0^2 . That is, consider a proper prior for ω that is a normal with mean ω_0 and variance σ_0^2 , i.e. $\omega \sim N(\omega_0, \sigma_0^2)$. Now since the normal is a proper conjugate prior family, extending the answer to problem II-B slightly, the posterior distribution for ω given N IID observations from a $N(\omega^*, \sigma^2)$ distribution (where σ^2 is a known variance), is $N(\omega(X), \sigma^2(X))$ where

$$\sigma^2(X) = \left(\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \right)^{-1}, \quad (19)$$

and $\omega(X)$ is

$$\omega(X) = \sigma^2(X) \left(\frac{\omega_0}{\sigma_0^2} + \frac{N\bar{X}_N}{\sigma^2} \right), \quad (20)$$

where again \bar{X}_N is the sample mean of $X = (x_1, \dots, x_N)$. Now as $\sigma_0^2 \rightarrow \infty$, stretching out the proper normal prior into an improper flat prior, the posterior mean converges to

$$\begin{aligned} \lim_{\sigma_0^2 \rightarrow \infty} \sigma^2(X) &= \frac{\sigma^2}{N} \\ \lim_{\sigma_0^2 \rightarrow \infty} \omega(X) &= \bar{X}_N, \end{aligned} \quad (21)$$

so by either approach the posterior distribution for normal model with unknown mean ω^* and known variance σ^2 when you are using an improper flat prior over the entire real line is a $N(\bar{X}_N, \frac{\sigma^2}{N})$ distribution. Clearly as $N \rightarrow \infty$ this proper posterior distribution converges in distribution to a unit mass on ω^* .

A final way to see the answer is to construct a proper prior over the interval $[-K, K]$. This prior has value $p(\omega) = \frac{1}{2K}$ and converges to 0 as $K \rightarrow \infty$. *However it not hard to show that for any fixed K the posterior distribution in this case is no longer normally distributed.* In particular, the support of the posterior must be a subset of the support of the prior distribution, which is the interval $[-K, K]$, as we showed above in problem III above. Since the support of a normal distribution is the entire real line, the posterior distribution in this case cannot be a normal distribution. It is not hard, however, to show the posterior is a *truncated normal distribution*. Further, as $K \rightarrow \infty$, the support expands to the entire real line, and posterior converges in distribution to $N(\bar{X}_N, \frac{\sigma^2}{N})$ just as in the case of limiting proper Normal prior derived above.

V. Consider the normal regression model

$$y_i = X_i\beta + \varepsilon_i \quad (22)$$

where $\varepsilon \sim N(0, \sigma^2)$ and X_i is $k \times 1$ vector that is a draw from some joint distribution $F(X)$ with density $f(X)$ that does not depend on β or σ^2 .

- A. Write the *full likelihood function* for the observations (y_i, X_i) , $i = 1, \dots, N$. Show that maximizing this likelihood $\mathcal{L}(y_1, X_1, \dots, y_N, X_N | \beta, \sigma^2)$ over the unknown parameters $\theta = (\beta, \sigma^2)$ is the same as the result of maximizing the *partial likelihood function* given by

$$\mathcal{L}_p(y_1, \dots, y_N | \theta, X_1, \dots, X_N) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma}} \exp\{-(y_i - X_i\beta)^2 / 2\sigma^2\} \quad (23)$$

Answer Let the density for the X covariates be $f(X)$. We assume that f does not depend on the unknown parameters of interest, $\theta = (\beta, \sigma^2)$. The full likelihood is then

$$\mathcal{L}(\theta) = \prod_{i=1}^N f(y_i | X_i, \theta) f(X_i) = \left(\frac{1}{\sqrt{2\pi\sigma}} \right)^N \left(\prod_{i=1}^N f(X_i) \right) \left(\prod_{i=1}^N \exp\{-(y_i - X_i\beta)^2 / 2\sigma^2\} \right) \quad (24)$$

and so it is evident that the part of the likelihood corresponding to the likelihood for the (X_1, \dots, X_N) 's is just a product of their marginal densities and factors out as a positive multiplicative constant that does not affect the result of the maximization over $\theta = (\beta, \sigma^2)$. Thus, in this case maximizing the partial likelihood over θ gives the same result as maximizing the full likelihood over θ and so there is no loss of information from ignoring the marginal density of the X_i observations. However if $f(X_i)$ did depend on θ this would no longer be true and there would be a loss of information from ignoring the part of the likelihood for the $f(X_i | \theta)$ densities, and maximizing the full and partial likelihood functions will no longer result in the identical values of $\hat{\theta}$. However, *extra credit*, can you show that maximizing the partial likelihood function still results in a consistent (but less efficient) estimator of θ ?

- B. Suppose you have a prior density $\pi(\theta)$ for $\theta = (\beta, \sigma^2)$. Show that the posterior density $\pi(\theta | y_1, X_1, \dots, y_N, X_N)$ does not depend on the density $f(X)$ of the X covariates.

Answer Just observe that if $f(X)$ does not depend on the parameters, it factors out of the likelihood function for the (y_i, X_i) data as a multiplicative constant and therefore cancels out of the numerator and denominator in the equation for the posterior density in (1).

- C. Suppose you have a non-informative prior over β and the *logarithm* of σ^2 . Specifically assume that the joint prior $\pi(\beta, \log(\sigma^2)) \propto \frac{1}{\sigma^2}$. Show that the posterior distribution for β conditional on σ^2 and $(y_1, X_1, \dots, y_N, X_N)$ is a multivariate Normal distribution with mean $\hat{\beta}$ and covariance matrix $\hat{\Sigma}$ given by

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'y \\ \hat{\Sigma} &= \sigma^2(X'X)^{-1}, \end{aligned}$$

where X is the $N \times k$ matrix of independent variables

$$X = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,k} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,k} \\ \cdots & \cdot & \cdot & \cdots \\ X_{N,1} & X_{N,2} & \cdots & X_{N,k} \end{bmatrix} \quad (25)$$

and y is the $N \times k$ vector of dependent variables in the regression

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{bmatrix}. \quad (26)$$

Answer This is a generalization of the answer to problem IV above. An answer (and full derivation of it) to this can be found in Edward Leamer's classic book, *Specification Searches: Ad Hoc Inference with Non Experimental Data* John Wiley and Sons, Inc., 1978. Alternatively the answer can be found online at <http://www.biostat.umn.edu/~sudiptob/ph8472/BayesianLinearModel.pdf>. I will sketch the answer below for convenience. Once again the strategy is to show that when there is a flat prior over β (a $K \times 1$ vector now) that when we do the multivariate integration over the β 's in the denominator of the general formula for the posterior distribution in equation (1) above, the denominator does not "blow up" but instead has a finite value, and this value cancels out in both the numerator and denominator of the equation for the posterior distribution in equation (1), resulting in a posterior distribution for β that is a normal distribution with the mean and covariance matrix given above — the standard OLS regression formulas for the OLS estimator of β and its covariance matrix. The trick is to convert the formula for the integral over β of the density of the product of the $f(y_i|X_i, \beta, \sigma^2)$ terms in the denominator of the formula for the posterior distribution in equation (1) to re-write it as a factor that does not depend on β times the integral of multivariate normal distribution over β which will integrate to 1. Thus, we will show that we can write

$$\int_{\beta} \prod_{i=1}^N f(y_i|X_i, \beta, \sigma^2) f(X_i) d\beta = \left[\prod_{i=1}^N g(y_i, X_i, \sigma^2) f(X_i) \right] \int_{\beta} \phi(\beta|\hat{\beta}, \hat{\Sigma}) d\beta = \left[\prod_{i=1}^N g(y_i, X_i, \sigma^2) f(X_i) \right], \quad (27)$$

where $g(y_i, X_i, \sigma^2)$ is a function of (y_i, X_i) that does not depend on β (which we will derive below) and $\phi(\beta|\hat{\beta}, \hat{\Sigma})$ is a multivariate normal density for β with mean $\hat{\beta}$ and $\hat{\Sigma}$, where the formulas for these are the OLS formulas given in equation (25) above. Note that since ϕ is a multivariate normal density, the integral of this density over all $\beta \in R^K$ equals 1, a fact we have used in equation (27) above. To derive the decomposition in equation (27), note that since $f(y_i|X_i, \beta, \sigma^2)$ is multivariate normal, we can use equation (23) above to write

$$\prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma}} \exp\{-(y_i - X_i\beta)^2/2\sigma^2\} = [2\pi\sigma^2]^{-N/2} \exp\{-\sum_{i=1}^N (y_i - X_i\beta)^2/2\sigma^2\}. \quad (28)$$

Note that we can write the latter exponential term using matrix (inner product) notation

$$\exp\{-\sum_{i=1}^N (y_i - X_i\beta)^2/2\sigma^2\} = \exp\{-\langle (y - X\beta), (y - X\beta) \rangle / (2\sigma^2)\} = \exp\{-(y - X\beta)'(y - X\beta)/2\sigma^2\}, \quad (29)$$

where y is the $N \times 1$ vector of all the dependent variables y_i and X is the $N \times K$ matrix of independent variables. Multiplying out the inner product inside this exponential we get

$$(y - X\beta)'(y - X\beta)/(2\sigma^2) = (y'y - 2y'X\beta + \beta'(X'X)\beta)/(2\sigma^2). \quad (30)$$

Now we use the trick of "completing the square" to rewrite this a function of (y, X) (and not β) plus the quadratic form $(\beta - \hat{\beta})'[\hat{\Sigma}^{-1}](\beta - \hat{\beta})$. Notice that when we exponentiate this quadratic form,

and multiply by a constant term $[(2\pi)^{-K/2}][|\hat{\Sigma}|^{-1/2}]$, the result is a multivariate normal distribution for β , and this implies that when we integrate over β in the denominator of the equation for the posterior distribution for β , the integral over β equals a production of functions that depend on the data and σ^2 but not β , time the integral of a multivariate normal density for β whose integral over all β is 1. Cancelling the factor of proportionality that does not depend on β from the numerator and denominator of the expression for the posterior distribution for β , (1), it follows that the posterior distribution for β is a multivariate normal density with mean $\hat{\beta}$ and covariance matrix $\hat{\Sigma}$.

$$\phi(\beta|\hat{\beta}, \hat{\Sigma}) = [(2\pi)^{-K/2}][|\hat{\Sigma}|^{-1/2}] \exp\{-(\beta - \hat{\beta})'[\hat{\Sigma}^{-1}](\beta - \hat{\beta})\}, \quad (31)$$

We leave it to you to complete the remaining details of showing how to complete the square and derive the $g(y_i, X_i, \sigma^2)$ function in equation (27) above. Note that this result is a generalization of the result we derived above for an improper prior for the parameter ω , the unknown mean of a normal data generating process for *IID* data (X_1, \dots, X_N) in problem IV above. *Extra credit*: show that the answer to problem IV is a special case of the answer just derived here.

- D. Show that the marginal posterior distribution for σ^2 is a scaled inverse Chi-squared distribution with $N - k$ degree of freedom and scale parameter \hat{s}^2 where

$$\hat{s}^2 = \frac{1}{N - k} (y - X'\hat{\beta})'(y - X'B) \quad (32)$$

Answer See the references cited in the answer to part C above.

VI. Read the paper, “Are People Bayesian? Uncovering Behavioral Strategies” by Mahmoud El-Gamal and David M. Grether in the *Journal of the American Statistical Association* (1995) **90-432** 1‘137–1145.

- A. The paper describes an experiment with human subjects where subjects were given a prior over which of two urns A and B (each containing 6 balls labelled ‘N’ and ‘G’) would be used to draw 6 balls that were shown to the subjects. The subjects’ task was to predict which of the two urns the 6 balls were drawn from. Urn A contained 4 N balls and 2 G balls, whereas urn B contained 3 N balls and 3 G balls. Suppose that the six balls were drawn *without replacement*. Show that a Bayesian decision maker would have a degenerate posterior probability that the balls were drawn from urns A and B (i.e. their posterior will be either 1 or 0).

Answer Since there are only 6 balls in each of two urns, drawing all of the balls out of one of these urns and revealing them to the subjects is tantamount to simply revealing which of the two different urns the balls were drawn from. So if the subject sees 4 N and 2 G balls, he/she concludes that the balls were drawn from urn A and has a posterior probability of 1 for urn A and 0 for urn B. If the subject sees 3 N and 3 G balls, he/she concludes that the balls were drawn from urn B and so has a posterior probability of 1 for urn B and 0 for urn A.

- B. Suppose that the 6 balls are drawn *with replacement*. Suppose that the sample that is drawn contains 3 N balls and 3 G balls and the prior probability of drawing from urn A as 1/6. What are the posterior probabilities that this sample was drawn from urns A and B?

Answer When the 6 balls are drawn with replacement it is no longer possible to know for certain which of the urns the sample was drawn from. But one can calculate the posterior probability using Bayes

Rule, a simple extension of the basic formula (1) above. For example, a sample containing 4 N and 2 G balls could have been drawn from either urn A or B when sampling is done with replacement, so this information does not allow us to definitively determine whether the balls were drawn from A or B. However the likelihood of this sample being drawn from urn A is higher than for urn B so a Bayesian would rationally adjust his/her prior probability of $\pi = 1/6$ to something higher after observing this sample. When sampling is done with replacement, the samples are *i.i.d.* Bernoulli draws where an N ball is drawn with probability p_A from urn A and probability p_B from urn B. For this case $p_A = 4/6 = 2/3$ and $p_B = 3/6 = 1/2$. Let the outcome Bernoulli variable be $B_i = 1$ if an N ball is the outcome of draw i from the urn, $i = 1, \dots, 6$. Then the likelihood a sample (B_1, \dots, B_6) given it was drawn from urn A is $\mathcal{L}(B_1, \dots, B_6|A)$

$$\mathcal{L}(B_1, \dots, B_6|A) = \prod_{i=1}^6 \left[\frac{2}{3} \right]^{B_i} \left[\frac{1}{3} \right]^{1-B_i} \quad (33)$$

and the likelihood for urn B is $\mathcal{L}(B_1, \dots, B_6|B)$

$$\mathcal{L}(B_1, \dots, B_6|B) = \prod_{i=1}^6 \left[\frac{1}{2} \right]^{B_i} \left[\frac{1}{2} \right]^{1-B_i} = \frac{1}{64} \quad (34)$$

Note that the likelihood of any sample is the same for urn B due to the equal number of N and G balls in that urn. Note also that $N = \sum_{i=1}^6 B_i$, the number of N balls in the sample of 6, is a *sufficient statistic* in the sense that the likelihood can be written as depending only on N (of course, the number of G balls will be $6 - N$). Thus we can write the likelihoods as $\mathcal{L}(N|A) = \left[\frac{2}{3} \right]^N \left[\frac{1}{3} \right]^{6-N}$ and $\mathcal{L}(N|B) = \frac{1}{64}$ and the posterior probability as $\pi(A|N)$ given by

$$\pi(A|N) = \frac{\frac{1}{6} \left[\left(\frac{2}{3} \right)^N \left(\frac{1}{3} \right)^{6-N} \right]}{\frac{1}{6} \left[\left(\frac{2}{3} \right)^N \left(\frac{1}{3} \right)^{6-N} \right] + \frac{5}{6} \left[\frac{1}{64} \right]}. \quad (35)$$

For a sample of $N = 3$ N balls and $6 - N = 3$ G balls, we can calculate $\pi(A|N = 3) = 0.1232$. Note that the prior probability of urn A being chosen is $\pi(A) = 1/6 \simeq 0.1667$, so the observation of a sample with an equal number of N and G balls reduces our belief that the chosen urn was urn A and increases our posterior probability that the balls were drawn from urn B from $\pi(B) = 5/6 \simeq .8333$ to $\pi(B|N = 3) = 1 - \pi(A|N = 3) = 0.8768$.

- C. Suppose the subject receives a reward of \$10 for a correct prediction of which urn the balls were drawn from, and \$0 for an incorrect prediction. If the subject was a Bayesian decision maker and an expected reward maximizer, describe the subject's *optimal decision rule* for choosing between the two urns. For each possible outcome with a prior probability that the draws were from urn A, provide the optimal prediction that maximizes the subject's expected rewards.

Answer If the subject is an expected reward maximizer and a Bayesian decision maker, the subject will choose the urn that has the highest expected monetary payoff. The expected payoff for choosing urn A is $10\pi(A|N)$ and the expected payoff for choosing urn B is $10\pi(B|N) = 10 - 10\pi(A|N)$. Thus, the subject will choose the urn that has the higher posterior probability of being the urn from which the balls were drawn, and will be indifferent if the posterior probabilities of the two urns are the same, $\pi(A|N) = 0.5 = \pi(B|N)$. In part B, the optimal choice for the subject would be to say that with a sample of $N = 3$ N balls and 3 G balls, the urn to report as the chosen urn would be urn B since it has a substantially higher posterior probability of being the correct urn.

D. Note that while this paper is investigating the question of whether *people are Bayesian*, the method of inference used by El-Gamal and Grether is a *classical method of inference*. Describe how this paper could have been analyzed from a Bayesian perspective if they believed that subjects were using one of K possible decision rules (including the optimal Bayesian decision rule in part C above) and their prior probabilities that any particular subject uses one of these decision rules is (p_1, \dots, p_K) . If they were willing to do this, would it have been possible for El-Gamal and Grether to use Bayesian methods to find the posterior probability distribution that any given subject is a Bayesian decision maker? Why or why not?

Answer The answer is yes, but this is not the approach to inference that Grether and El-Gamal used in their paper. Instead they used a *classical* (i.e. non-Bayesian) approach called the *Estimation-Classification Algorithm* (*EC* algorithm) that for each subject evaluates the likelihood of their sequence of choices under each of the possible decision rules, $1, \dots, K$ and chooses as the decision rule the one with the highest likelihood. They then estimated the overall fraction of students whose reporting behavior was best described by the Bayesian decision rule. Note that if subjects are given repeated urn choice questions, the likelihood of behaving according to the Bayesian decision rule will be equal to 0 unless the subject reports exactly as prescribed by the Bayesian decision rule (i.e. reporting the urn with the highest posterior probability of being the urn) in each single instance. Since it is unlikely to be the case that any subject will behave exactly in a Bayesian way in *every single trial they are asked to make a decision on* one approach is to extend the model to add some sort of “reporting error” where the subject who is a Bayesian may with some small probability report “incorrectly” i.e. report their guess of the urn that was used to draw the sample of 6 balls as one that is not the one with the highest posterior probability. There are deeper philosophical issues posed by this strategy about whether a subject can be described as a Bayesian if they occasionally make mistakes and report the urn that has a lower posterior probability of being the correct urn. But regardless, Grether and El-Gamal found that a substantial fraction of students could not be described as Bayesian decision makers, even when some “reporting error” is allowed for, and if we were to adopt a strict definition and not allow the possibility of reporting error, then only a small minority of student respondents could be described as “perfect Bayesian decision makers”. If this is true, what does this suggest about the relevance of Bayesian econometrics and Bayesian models of decision making in economics more generally (i.e. the frequently used concept of “Bayesian Nash equilibrium” where players in a game are assumed to be behaving according to Bayes rule and make use of all available information in the game to calculate the posterior probability distribution over the “types” of their opponents given the history of the at each possible point)?

VII. Read sections 1 and 2 of “Introduction to Statistical Learning Theory” by Bosquet, Boucheron and Lugosi.

Answer Everyone is assumed to have done this and we give points for this as a free bonus.