Homework Five: Solutions
Dancsi Percival, Assisting the Teacher in a Friendly Way
Much credit also to Zachary Kurtz, an antiquated TA on whose solutions these draw.

1 (a) Let’s talk about this problem. Instead of doing an integral, we can take advantage of properties of the normal distributions and play tricks. So, since the likelihood \( Y|\beta, M \sim \mathcal{N}(X\beta, I) \), we can write:

\[
Y = X\beta + \epsilon, \quad \text{with} \quad \epsilon|M \sim \mathcal{N}(0, I),
\]

and we know from the prior:

\[
\beta|M \sim \mathcal{N}(\mu_\beta, \Sigma_\beta).
\]

Thus, to get at the marginal distribution of \( Y \), that is with \( \beta \) allowed to be random, we just apply the properties that we know about normals to obtain:

\[
Y|M \sim \mathcal{N}(X\mu_\beta, X\Sigma_\beta X^T + I) \quad (1)
\]

(b) Use your proportionality:

\[
p(\beta|Y, M) \propto p(Y|\beta, M)p(\beta|M) \quad (2)
\]

\[
\propto \exp \left( -\frac{1}{2} \left[ (Y - X\beta)^T(Y - X\beta) + (\beta - \mu_\beta)^T \Sigma_\beta^{-1}(\beta - \mu_\beta) \right] \right) \quad (3)
\]

\[
\propto \exp \left( -\frac{1}{2} \beta^T X^T X \beta - 2\beta^T X^T Y + \beta^T \Sigma_\beta^{-1} \beta - 2\beta^T \Sigma_\beta^{-1} \mu_\beta \right) \quad (4)
\]

So, this is the kernel of a normal distribution, with parameters:

\[
\Sigma_* = (X^T X + \Sigma_\beta^{-1})^{-1} \quad (5)
\]

\[
\mu_* = \Sigma_* (X^T Y + \Sigma_\beta^{-1} \mu_\beta) \quad (6)
\]

How to see this: looking at the \( \beta^T \beta \) terms, the middle of that tells us what \( \Sigma_*^{-1} \) is. Then, whatever is next to the \( -2\beta^T \) term is \( \Sigma_*^{-1} \mu_* \).

(c) What is going on here is just a simple check of Bayes' rule, since:

\[
p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} \quad (7)
\]

\[
= \frac{p(Y, \beta)}{p(Y)} \quad (8)
\]

\[
\implies p(Y) = \frac{p(Y|\theta)p(\theta)}{p(\theta|Y)} \quad (9)
\]

Just a quick check:

(c) Methods: Ratio calculation, integral calculation (using random beta)
[1] -157.9532
[1] -157.9532
These look the same! (these are on log scale)
(c) Methods: Ratio calculation, integral calculation (using random beta)
[1] 2.522181e-69
[1] 2.522181e-69
(d, e) Here we are using a simple monte carlo integration technique. Ignoring all the technical details:

\[ E_f(g(x)) = \int g(x)f(x)dx \]  

(10)

So to approximate this integral we can, (1) sample many \( x \) from the density \( f(x) \), (2) calculate \( \frac{1}{n} \sum_i g(x_i) \). In our problem, we sample from the (d) prior / (e) posterior, and then return the average (d) likelihood (e) one over the likelihood of these samples:

Mean value of methods: prior sampling, posterior sampling
[1] 1.888512e-69
[1] 3.346707e-66

Variance of methods: prior sampling, posterior sampling
[1] 4.882828e-137
[1] 2.610803e-132

(f) I'd use the prior sampling scheme!

2 Some general comments about problems two and three. For (a), the marginal likelihood, given a model, is \( P(Y|M) \) — here "marginal" means marginal with respect to the parameters \( (\beta) \). In (a), we know this distribution exactly from problem 1.

In (b), the posterior model probability is \( P(M|Y) \), by Bayes’ rule:

\[ P(M|Y) = \frac{P(Y|M)P(M)}{P(Y)} \]  

(11)

where, \( P(M) \) is the model prior, given to us in both problems (and in general known since we set it!).

Notice that the denominator is now another marginal distribution \( P(Y) \), which is the marginal with respect to models and parameters now, by the law of total probability:

\[ P(Y) = \sum_i P(Y|M_i)P(M_i) \]  

(12)

(c) is asking the posterior probability of seeing some event in \( \beta \), call it \( \Delta_\beta \), given the data, this may be a bit abstract, but you can use the law of total probability here:

\[ P(\Delta_\beta|Y) = \sum_i P(\Delta_\beta|Y,M_i)P(M_i|Y) \]  

(13)

The second term in that summation is the posterior probability that we obtained in (b). The first term is simply a probability calculation under a certain model, which is a feature of the posterior distribution \( p(\beta|Y,M) \), which we known the form of from problem 1.

(a-d) See the code.
DATA SET 1:
MODEL 1 \( P(Y|M) = 1.404174 \times 10^{-67} \) (log(P(Y|M)) = -153.9338 )
MODEL 2 \( P(Y|M) = 7.516445 \times 10^{-68} \) (log(P(Y|M)) = -154.5587 )
MODEL 3 \( P(Y|M) = 4.712609 \times 10^{-68} \) (log(P(Y|M)) = -155.0255 )
MODEL 4 \( P(Y|M) = 3.287053 \times 10^{-68} \) (log(P(Y|M)) = -155.3858 )
MODEL 5 \( P(Y|M) = 2.455264 \times 10^{-68} \) (log(P(Y|M)) = -155.6776 )

DATA SET 2:
MODEL 1 \( P(Y|M) = 1.296372 \times 10^{-72} \) (log(P(Y|M)) = -165.5266 )
MODEL 2 \( P(Y|M) = 2.290943 \times 10^{-68} \) (log(P(Y|M)) = -155.7468 )
MODEL 3 \( P(Y|M) = 5.223243 \times 10^{-67} \) (log(P(Y|M)) = -152.6201 )
MODEL 4 \( P(Y|M) = 2.25142 \times 10^{-66} \) (log(P(Y|M)) = -151.1591 )
MODEL 5 \( P(Y|M) = 5.055798 \times 10^{-66} \) (log(P(Y|M)) = -150.3501 )

(b) [posterior model probabilities P(M|Y) rounded off]

<table>
<thead>
<tr>
<th></th>
<th>Data Set 1</th>
<th>Data Set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.439</td>
<td>0.000</td>
</tr>
<tr>
<td>Model 2</td>
<td>0.235</td>
<td>0.003</td>
</tr>
<tr>
<td>Model 3</td>
<td>0.147</td>
<td>0.067</td>
</tr>
<tr>
<td>Model 4</td>
<td>0.103</td>
<td>0.287</td>
</tr>
<tr>
<td>Model 5</td>
<td>0.077</td>
<td>0.644</td>
</tr>
</tbody>
</table>

(c)

| P(B_1 < 0 | Y):          | Data Set 1 | Data Set 2 |
|------------|------------|------------|
| Model 1    | 0.1789310  | 1          |
| Model 2    | 0.1998018  | 1          |
| Model 3    | 0.2072609  | 1          |
| Model 4    | 0.2110883  | 1          |
| Model 5    | 0.2134165  | 1          |

Model averaged probabilities for P(B_1 | Y):

<table>
<thead>
<tr>
<th></th>
<th>Data Set 1</th>
<th>Data Set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1939484</td>
<td>1.0000000</td>
</tr>
</tbody>
</table>

3

[1] "The models we consider are..."
I(Variable 1 neq 0) I(Variable 2 neq 0) I(Variable 3 neq 0)

<table>
<thead>
<tr>
<th></th>
<th>I(Variable 1)</th>
<th>I(Variable 2)</th>
<th>I(Variable 3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Model 2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Model 3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Model 4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(a)

Marginal likelihood, variable selection

<table>
<thead>
<tr>
<th></th>
<th>Data Set 1</th>
<th>Data Set 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>1.031442e-71</td>
<td>1.802741e-67</td>
</tr>
<tr>
<td>Model 2</td>
<td>2.600575e-70</td>
<td>4.121314e-265</td>
</tr>
</tbody>
</table>
Here are some general comments about importance sampling. A nice technique is Monte Carlo integration. Suppose we wish to calculate:

$$E_f(g(x)) = \int g(x)f(x)dx \quad (14)$$

So to approximate this integral we can, (1) sample many $x$ from the density $f(x)$, (2) calculate $\frac{1}{n} \sum_i g(x_i)$.

Now, here a simple trick we can play, multiply and divide by another density, $h(x)$:

$$E_f(g(x)) = \int g(x)f(x)\frac{h(x)}{h(x)}dx = E_h \left( g(x)\frac{f(x)}{h(x)} \right) \quad (15)$$

The second line is a simple property of expectations. Now, we simply do this integral using our Monte Carlo integration strategy, where $h(x)$ is the density of $x$, and $g^*(x) = g(x)\frac{f(x)}{h(x)}$. So we, (1) sample many $x$ from the density $h(x)$, (2) calculate $\frac{1}{n} \sum_i g(x_i)\frac{f(x)}{h(x)}$. That is all there is to importance sampling.

Problem 4

"TRUE PROBABILITY:"

0.06966298

"DIRECT DRAW:"

0.069414 (mean)

0.007834666 (variance)

"MEANS"

SHIFT = 0.5  SHIFT = 1  SHIFT = 1.5  SHIFT = 2  SHIFT = 2.5  SHIFT = 3

0.06966298  0.06966298  0.06966298  0.06966298  0.06966298  0.06966298

"VARIANCES"

SHIFT = 0.5  SHIFT = 1  SHIFT = 1.5  SHIFT = 2  SHIFT = 2.5  SHIFT = 3

0.007834666  0.007834666  0.007834666  0.007834666  0.007834666  0.007834666

The best shift? 2.5 shift.

"Every reader, as he reads, is actually the reader of himself."