**Solutions: Assignment #3 STA355H1S**

1. (a) The posterior density of $(\alpha, \lambda)$ is proportional to

$$\pi(\alpha,\lambda)\mathcal{L}(\alpha,\lambda) = \frac{1}{10000}\exp\left(-\frac{\alpha}{100} - \frac{\lambda}{100}\right)\frac{\lambda^{n\alpha}(x_1 \times \cdots \times x_n)^{\alpha-1}\exp\left(-\lambda\sum_{i=1}^n x_i\right)}{[\Gamma(\alpha)]^n}$$

Factoring out all terms not involving $\lambda$, we're left with the following integral:

$$\int_0^\infty \lambda^{n\alpha}\exp\left[-\lambda\left(\sum_{i=1}^n x_i + \frac{1}{100}\right)\right]\,d\lambda.$$

Making the change of variables

$$u = \lambda\left(\sum_{i=1}^n x_i + \frac{1}{100}\right) \quad \text{with} \quad du = d\lambda\left(\sum_{i=1}^n x_i + \frac{1}{100}\right)$$

and so the integral above equals

$$\left(\sum_{i=1}^n x_i + \frac{1}{100}\right)^{-(n\alpha+1)}\Gamma(n\alpha+1).$$

(b) We need to compute

$$\int_0^\infty \frac{\Gamma(n\alpha+1)}{[\Gamma(\alpha)]^n}\exp\left(\alpha\sum_{i=1}^n \ln(x_i) - \alpha/100\right)\left(\frac{1}{100} + \sum_{i=1}^n x_i\right)^{-n\alpha-1}\,d\alpha.$$

This is complicated by the fact that $\Gamma(x)$ cannot be computed in R (using the function `gamma`) for larger $x$ ($x > 171$). To do the integration, we take the logarithm of the integrand (which is computable) and then subtract its maximum value; in this case, we can use the function `lgamma`, which computes the logarithm of the Gamma function. Converting back to the original scale, we now have an integrand taking values between 0 and 1, which can be easily integrated numerically. The following R code does this pre-normalization:

```
> x <- scan("aircon.txt")
> n <- length(x)
> alpha <- c(6000:12000)/10000
> logint <- lgamma(n*alpha+1) - n*lgamma(alpha)
> logint <- logint + alpha*sum(log(x)) - alpha/100
> logint <- logint - (n*alpha-1)*log(sum(x)+1/100)
> logint <- logint - max(logint)
> int <- exp(logint)
```

We can now use the following R code to compute the normalizing constant:

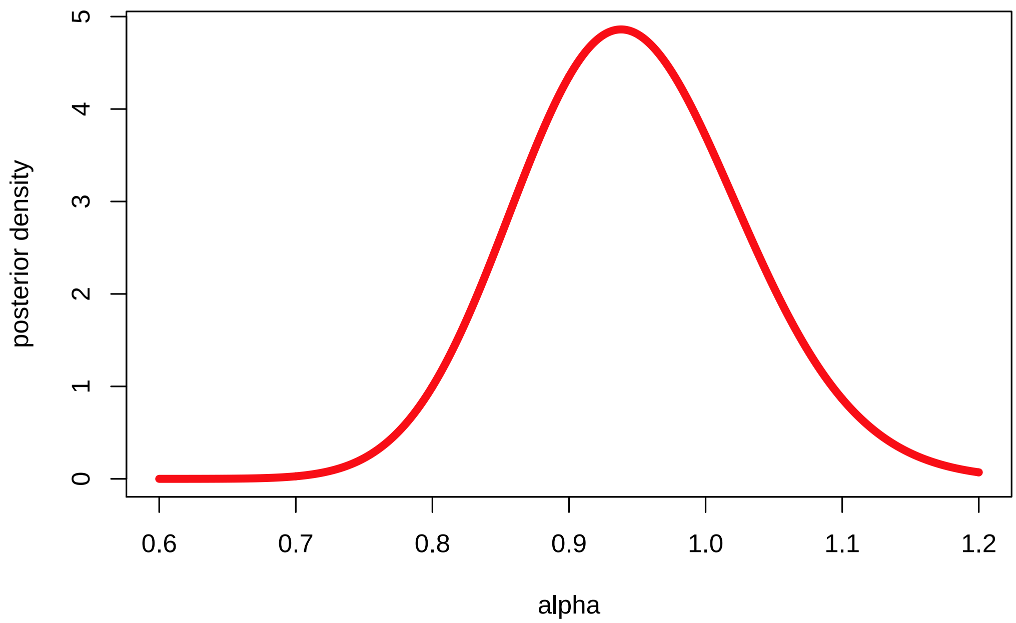


Figure 1: Posterior density of $\alpha$ for air conditioning data.

```
> mult <- c(1/2,rep(1,5999),1/2)/10000
> norm <- sum(mult*int)
> post <- int/norm
> plot(alpha,post,type="l",ylab="posterior density",lwd=3,col="red")
```

The posterior density is shown in Figure 1.

(c) This is much (much!) easier than it seems! We have

$$P(\alpha = 1|x_1, \cdots, x_n) = \frac{\displaystyle\int_0^\infty \pi(\lambda, 1)\mathcal{L}(\lambda, 1)\, d\lambda}{\displaystyle\int_0^\infty \pi(\lambda, 1)\mathcal{L}(\lambda, 1)\, d\lambda + \int_0^\infty \int_0^\infty \pi(\lambda, \alpha)\mathcal{L}(\lambda, \alpha)\, d\lambda\, d\alpha}$$

If we divide the numerator and denominator by $\int_0^\infty \int_0^\infty \pi(\lambda, \alpha)\mathcal{L}(\lambda, \alpha)\, d\lambda\, d\alpha$ then we have

$$P(\alpha = 1|x_1, \cdots, x_n) = \frac{\theta\pi(1|x_1, \cdots, x_n)}{\theta\pi(1|x_1, \cdots, x_n) + 1 - \theta}$$

where $\pi(1|x_1, \cdots, x_n)$ was computed in part (b). Table 1 below gives the posterior probabilities; note that the posterior probability is greater than the prior probability $\theta$. This suggests that the Exponential model is probably a reasonably good approximation.

| $\theta$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| $P(\alpha = 1\vert\text{data})$ | 0.29 | 0.48 | 0.61 | 0.71 | 0.79 | 0.85 | 0.90 | 0.94 | 0.97 |

Table 1: Posterior probabilities of Exponential model as a function of the prior probability $\theta$.

2. (a) We use the cross-validation (CV) estimate from Assignment 2, which has its mode at $x = 0.07946$. (The default bandwidth for `density` has its mode at $x = 0.07744$.) Table 2 below gives the Venter estimates for various values of $\tau$. Note that for smaller values of $\tau$, the Venter estimates are close to the mode of the CV estimate while for larger $\tau$, the Venter estimates are close to the mode of the default density estimate.

| $\tau$ | 0.05 | 0.10 | 0.15 | 0.20 | 0.25 | 0.30 | 0.35 | 0.40 | 0.45 | 0.50 |
|---|---|---|---|---|---|---|---|---|---|---|
| Venter est. | 0.0793 | 0.0794 | 0.0796 | 0.0794 | 0.0790 | 0.0787 | 0.0780 | 0.0761 | 0.0757 | 0.0754 |

Table 2: Venter estimates for different values of $\tau$.

(b) The following function can be used to do the simulations for various values of $n$ and $\alpha$:

```
ventermse <- function(n,tau,alpha,nrep=10000) {
               if (missing(tau)) tau <- c(1:10)/20
               modest <- NULL
               for (i in 1:nrep) {
                  m <- NULL
                  x <- rgamma(n,alpha)
                  for (i in tau) {
                     m <- c(m,venter(x,tau=i))
                     }
                  modest <- rbind(modest,m)
                  }
               bias <- apply(modest-(alpha-1),2,mean)
               variance <- apply(modest,2,var)
               mse <- apply((modest-(alpha-1))^2,2,mean)
               r <- list(modes=modest,mse=mse,bias=bias,var=variance)
               r
               }
```

We obtain the following results (note that your results will be different due to the randomness of the simulation):

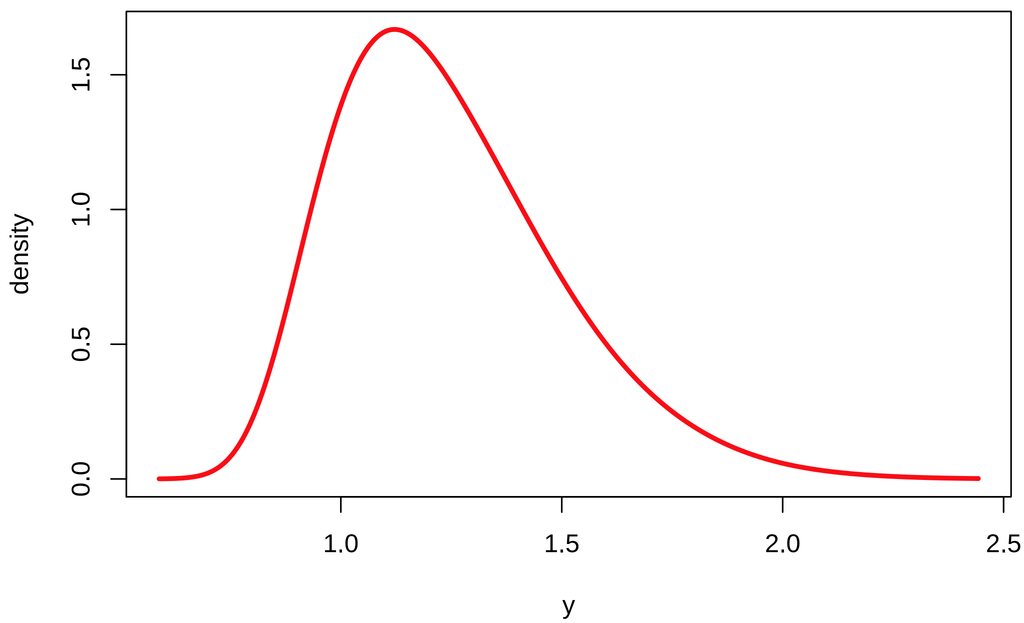


Figure 2: Estimated density of the Venter estimator for $\tau = 1/2$ and $n = 100$ for Gamma(2,1) distribution.

```
> r <- ventermse(100,tau=c(0.1,0.5),alpha=2)
> round(r$mse,4)
[1] 0.2672 0.1261
> r <- ventermse(1000,tau=c(0.1,0.5),alpha=2)
> round(r$mse,5)
[1] 0.07087 0.04751
> r <- ventermse(100,tau=c(0.1,0.5),alpha=10)
> round(r$mse,4)
[1] 1.8313 0.7139
> r <- ventermse(1000,tau=c(0.1,0.5),alpha=10)
> round(r$mse,4)
[1] 0.5479 0.1640
```

Note that the Venter estimator with $\tau = 0.5$ is better in terms of MSE for all four scenarios. A closer examination of the simulation data reveals that this superiority is due to its smaller variance; the estimator with $\tau = 0.1$ has a smaller (absolute) bias in all cases.

(c) We can use the following R code to compute an estimate of the density of the Venter estimator for $\alpha = 2$ and $n = 100$:

```
> muhat <- NULL
```

```
> sumx <- NULL
> for (i in 1:10000) {
+    x <- rgamma(100,2)
+    muhat <- c(muhat,venter(x,tau=1/2))
+    sumx <- c(sumx,sum(x))
+    }
> y <- min(muhat) + (max(muhat)-min(muhat))*c(0:5000)/5000
> fmuhat <- NULL
> for (z in y) {
+    fmuhat <- c(fmuhat,mean(dgamma(z,shape=200,scale=muhat/sumx)))
+    }
> plot(y,fmuhat,type="l",ylab="density",lwd=3,col="red")
```

The estimated density of the Venter estimator is shown in Figure 2; note that this density is skewed to the left.

**Supplemental problems:**

3. (a) Since $X_i$ is deleted from the sample, $\widehat{\theta}_{-i}$ satisfies the equation

$$\sum_{j \neq i} \ell'(X_j; \widehat{\theta}_{-i}) = 0.$$

Adding $\ell'(X_i; \widehat{\theta}_{-i})$ to both sides of the equation, we get

$$\sum_{j=1}^{n} \ell'(X_j; \widehat{\theta}_{-i}) = \ell'(X_i; \widehat{\theta}_{-i})$$

(b) By a Taylor series approximation,

$$\ell'(X_j; \widehat{\theta}_{-i}) \approx \ell'(X_j; \widehat{\theta}) + (\widehat{\theta}_{-i} - \widehat{\theta})\ell''(X_j; \widehat{\theta})$$

Thus (assuming that the approximation above is good), we have

$$\begin{aligned}
\ell'(X_i; \widehat{\theta}_{-i}) &= \sum_{j=1}^{n} \ell'(X_j; \widehat{\theta}_{-i}) \\
&\approx \sum_{j=1}^{n} \ell'(X_j; \widehat{\theta}) + (\widehat{\theta}_{-i} - \widehat{\theta}) \sum_{j=1}^{n} \ell''(X_j; \widehat{\theta}) \\
&= (\widehat{\theta}_{-i} - \widehat{\theta}) \sum_{j=1}^{n} \ell''(X_j; \widehat{\theta})
\end{aligned}$$

since $\widehat{\theta}$ is the MLE (and so $\sum_{j=1}^{n} \ell'(X_j; \widehat{\theta}) = 0$). Thus

$$\widehat{\theta}_{-i} - \widehat{\theta} \approx \frac{\ell'(X_i; \widehat{\theta}_{-i})}{\sum_{j=1}^{n} \ell''(X_j; \widehat{\theta})} \approx \frac{\ell'(X_i; \widehat{\theta})}{\sum_{j=1}^{n} \ell''(X_j; \widehat{\theta})}$$

where the latter approximation follows from the continuity of $\ell'$ and the fact that $\widehat{\theta}_{-i} - \widehat{\theta}$ is not too large. Likewise (again having faith in our approximations!),

$$\begin{aligned}
\widehat{\theta}_{\bullet} &= \frac{1}{n} \sum_{i=1}^{n} \widehat{\theta}_{-i} \\
&\approx \widehat{\theta} + \frac{\sum_{i=1}^{n} \ell'(X_i; \widehat{\theta})}{\sum_{j=1}^{n} \ell''(X_j; \widehat{\theta})} \\
&= \widehat{\theta}
\end{aligned}$$

since (again) $\widehat{\theta}$ is the MLE.

(c) The jackknife standard error estimator is

$$\widehat{\mathrm{se}}_{\mathsf{jk}}(\widehat{\theta}) = \left\{ \frac{n-1}{n} \sum_{i=1}^{n} (\widehat{\theta}_{-i} - \widehat{\theta}_{\bullet})^2 \right\}^{1/2}$$

and using our approximations from part (b), we get

$$\widehat{\mathrm{se}}_{\mathsf{jk}}(\widehat{\theta}) \approx \left\{ \frac{n-1}{n} \frac{\sum_{i=1}^{n} [\ell'(X_i; \widehat{\theta})]^2}{\left[\sum_{j=1}^{n} \ell''(X_j; \widehat{\theta})\right]^2} \right\}^{1/2}$$

In constrast, the estimator based on the observed Fisher information is

$$\widehat{\mathrm{se}}_{\mathsf{fi}}(\widehat{\theta}) = \left\{ -\sum_{i=1}^{n} \ell''(X_i; \widehat{\theta}) \right\}^{1/2}$$

The difference between the two estimators is easily explained: The estimator based on observed Fisher information assumes that the random variables $X_1, \cdots, X_n$ come from the density or mass function $f(x; \theta)$; on the other hand, the jackknife estimator assumes non-parametric model where $X_1, \cdots, X_n$ come from an unknown distribution function $F$ and $\widehat{\theta}$ is a substitution principle estimator of $\theta = \theta(F)$ satisfying the equation

$$E_F[\ell'(X_i; \theta(F))] = 0$$

where $\ell'$ need not have any relation to the underlying distribution function $F$. However, in the case where $F$ has a density or mass function of the form $f(x; \theta(F))$ then (assuming maximum likelihood regularity conditions)

$$\mathrm{Var}_F[\ell'(X_i; \theta(F))] = -E_F[\ell''(X_i; \theta(F))]$$

and so it follows that

$$\frac{1}{n} \sum_{i=1}^{n} [\ell'(X_i; \widehat{\theta})]^2 \approx \mathrm{Var}_F[\ell'(X_i; \theta(F))] = -E_F[\ell''(X_i; \theta(F))] \approx -\frac{1}{n} \sum_{i=1}^{n} \ell''(X_i; \widehat{\theta}),$$

in which case the two standard error estimators should be quite similar.

(d) For the Exponential model, $\widehat{\lambda} = 1/\bar{X}$ and $\widehat{\mathrm{se}}_{\mathsf{fi}}(\widehat{\lambda}) = \widehat{\lambda}/\sqrt{n}$. For the air conditioning data, $\widehat{\lambda} = 0.01091$ and $\widehat{\mathrm{se}}_{\mathsf{fi}}(\widehat{\lambda}) = 0.01091/\sqrt{199} = 0.000773$. The jackknife estimate can be computed as follows:

```
> loo <- NULL
> for (i in 1:199) {
+   xi <- aircon[-i]
+   loo <- c(loo,1/mean(xi))
+   }
> jack.se <- sqrt(198*sum((loo-mean(loo))^2)/199)
> jack.se
[1] 0.0008897483
```

The two estimates are different but not too dissimilar. This is consistent with the analysis in Problem 1 where we observed that the Exponential model was "not too bad".

4. (a) First of all,

$$\begin{aligned}
I_X(\theta) &= \int_{-\infty}^{\infty} \left[\frac{\partial}{\partial\theta}\ln f(x;\theta)\right]^2 f(x;\theta)\,dx \\
&= \sum_{k=1}^{m}\int_{B_k} \left[\frac{\partial}{\partial\theta}\ln f(x;\theta)\right]^2 f(x;\theta)\,dx \\
&= \sum_{k=1}^{m}\int_{B_k} p(k;\theta)\left[\frac{\partial}{\partial\theta}\ln f(x;\theta)\right]^2 \frac{f(x;\theta)}{p(k;\theta)}\,dx
\end{aligned}$$

Next,

$$\begin{aligned}
I_Y(\theta) &= \sum_{k=1}^{m}\left[\frac{\partial}{\partial\theta}\ln p(k;\theta)\right]^2 p(k;\theta) \\
&= \sum_{k=1}^{m}\left[\frac{\frac{\partial}{\partial\theta}p(k;\theta)}{p(k;\theta)}\right]^2 p(k;\theta)
\end{aligned}$$

and

$$\begin{aligned}
\left[\frac{1}{p(k;\theta)}\frac{\partial}{\partial\theta}p(k;\theta)\right]^2 &= \left[\frac{1}{p(k;\theta)}\frac{\partial}{\partial\theta}\int_{B_k} f(x;\theta)\,dx\right]^2 \\
&= \left[\frac{1}{p(k;\theta)}\int_{B_k}\frac{\partial}{\partial\theta} f(x;\theta)\,dx\right]^2 \\
&= \left[\int_{B_k}\frac{\partial}{\partial\theta}\ln f(x;\theta)\frac{f(x;\theta)}{p(k;\theta)}\,dx\right]^2 \\
&\le \int_{B_k}\left[\frac{\partial}{\partial\theta}\ln f(x;\theta)\right]^2\frac{f(x;\theta)}{p(k;\theta)}\,dx
\end{aligned}$$

by Jensen's inequality since $f(x;\theta)/p(k;\theta)$ is a density on $B_k$ (it is the conditional density of $X_i$ given $X_i \in B_k$ or $Y_i = k$). Therefore,

$$\begin{aligned}
I_Y(\theta) &\le \sum_{k=1}^{m} p(k;\theta)\int_{B_k}\left[\frac{\partial}{\partial\theta}\ln f(x;\theta)\right]^2\frac{f(x;\theta)}{p(k;\theta)}\,dx \\
&= I_X(\theta).
\end{aligned}$$

(b) Note that $E(U^2) = [E(U)]^2$ if, and only if, $U$ is a constant with probability 1, and $E(U^2) \approx [E(U)]^2$ if $U$ is nearly constant, that is, $\mathrm{Var}(U) = E(U^2) - [E(U)]^2$ is very small. This suggests that $I_X(\theta) \approx I_Y(\theta)$ if for each $B_k$,

$$\frac{\partial}{\partial\theta}\ln f(x;\theta) \approx a_k(\theta) \quad \text{for } x \in B_k,$$

which occurs if $B_k$ is sufficiently small that for each $\theta$ $f(x;\theta)$ is nearly constant for $x \in B_k$. Thus the information loss due to "discretization" decreases as size of the sets $\{B_k\}$ become smaller.

As an aside, it is possible to construct examples where $I_X(\theta) = I_Y(\theta)$, although these examples are very artificial. Suppose that

$$f(x;\theta) = c_k(x)p(k;\theta) \quad \text{for } x \in B_k$$

where $c_k(x)$ is a density on $B_k$. Then

$$\frac{\partial}{\partial\theta} \ln f(x;\theta) = \frac{\partial}{\partial\theta} \ln p(k;\theta) \quad \text{for } x \in B_k$$

and following the steps above, we have $I_X(\theta) = I_Y(\theta)$. Intuitively this makes sense — given an observation $X$ from $f(x;\theta)$, knowing into which $B_k$ it falls tells us everything we need to know about $\theta$.

5. (a) The log-likelihood function is

$$\ln \mathcal{L}(\beta) = \sum_{i=1}^{n} \{-\beta(x_i - m_0 - \delta/2)\} + n\ln[1 - \exp(-\beta\delta)]$$

Differentiating and solving, we get obtain the MLE

$$\widehat{\beta} = \frac{1}{\delta} \ln\left(1 + \frac{\delta}{\bar{X} - m_0 - \delta/2}\right).$$

(b) We can compute the MLE using the code given below; the vector `mag` contains the magnitudes.

```
> delta <- 0.1
> m0 <- 4.95
> mle <- log(1 + delta/(mean(mag)-m0-delta/2))/delta
> mle
[1] 2.055946
```

The jackknife standard error estimate is computed as follows:

```
> mle.i <- NULL
> for (i in 1:433) {
+    mag.i <- mag[-i]
+    mle.i <- c(mle.i,log(1 + delta/(mean(mag.i)-m0-delta/2))/delta)
+    }
> jack.var <- 432*sum((mle.i-mean(mle.i))^2)/433
> jack.se <- sqrt(jack.var)
> jack.se
[1] 0.09787367
```

For the observed Fisher information, we have

$$-\frac{d^2}{d\beta^2}\ln\mathcal{L}(\widehat{\beta}) = \frac{n\delta^2\exp(-\widehat{\beta}\delta)}{(1-\exp(-\widehat{\beta}\delta))^2}$$

```
> fisher.info <- 433*delta^2*exp(-delta*mle)/(1-exp(-delta*mle))^2
> fisher.se <- 1/sqrt(fisher.info)
> fisher.se
[1] 0.09897655
```

The (approximate) 95% confidence intervals are $2.056 \pm 0.192$ (using the jackknife standard error estimate) and $2.056 \pm 0.194$ (using the standard error estimate based on the observed Fisher information). The two intervals are clearly very similar.

6. (a) The log-likelihood function is

$$\ln\mathcal{L}(\theta) = 2x_1\ln(\theta) + x_2\left\{\ln(2)+\ln(\theta)+\ln(1-\theta)\right\} + 2x_3\ln(1-\theta).$$

Differentiating with respect to $\theta$, we get

$$\frac{d}{d\theta}\ln\mathcal{L}(\theta) = \frac{(2x_1+x_2)(1-\theta)-(x_2+2x_3)\theta}{\theta(1-\theta)}$$

and so the MLE is

$$\widehat{\theta} = \frac{2X_1+X_2}{2(X_1+X_2+X_3)} = \frac{2X_1+X_2}{2n}.$$

The observed Fisher information is

$$-\frac{d^2}{d\theta^2}\ln\mathcal{L}(\widehat{\theta}) = \frac{2X_1+X_2}{\widehat{\theta}^2} + \frac{X_2+2X_3}{(1-\widehat{\theta})^2} = \frac{2n}{\widehat{\theta}(1-\widehat{\theta})}.$$

Thus the estimated standard error is

$$\widehat{\mathrm{se}}(\widehat{\theta}) = \left\{\frac{\widehat{\theta}(1-\widehat{\theta})}{2n}\right\}^{1/2}.$$

(b) Note that we can write the prior as

$$\pi(\theta) = K(\alpha,\beta)\,\theta^{\alpha-1}(1-\theta)^{\beta-1} \quad \text{for } 0\le\theta\le 1$$

where the normalizing constant is

$$K(\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

The posterior density is

$$\pi(\theta|x_1,x_2,x_3) = \frac{\pi(\theta)\mathcal{L}(\theta)}{\int_0^1 \pi(s)\mathcal{L}(s)\,ds};$$

note that

$$\pi(\theta)\mathcal{L}(\theta) = \text{constant} \times \theta^{\alpha+2x_1+x_2-1}(1-\theta)^{\beta+x_2+2x_3-1}$$

where the constant depends on $\alpha, \beta, x_1, x_2, x_3$ but not $\theta$. Therefore,

$$\pi(\theta|x_1, x_2, x_3) = K(\alpha + 2x_1 + x_2, \beta + x_2 + 2x_3)\, \theta^{\alpha+2x_1+x_2-1}(1-\theta)^{\beta+x_2+2x_3-1}$$

and so the posterior is also a Beta distribution with hyperparameters $\alpha + 2x_1 + x_2$ and $\beta + x_2 + 2x_3$.
(As an aside, the mean and variance of the posterior distribution are

$$\begin{aligned} E(\theta|x_1, x_2, x_3) &= \int_0^1 \theta\pi(\theta|x_1, x_2, x_3)\, d\theta \\ &= \frac{\alpha + 2x_1 + x_2}{\alpha + \beta + 2n} \\ \text{Var}(\theta|x_1, x_2, x_3) &= \frac{(\alpha + 2x_1 + x_2)(\beta + x_2 + 2x_3)}{(\alpha + \beta + 2n)^2(\alpha + \beta + 2n + 1)} \end{aligned}$$

If $2x_1+x_2$ and $x_2+2x_3$ are large compared to the prior hyperparameters $\alpha$ and $\beta$, respectively then the posterior mean and variance can be approximated $\widehat{\theta}$ and $\widehat{\theta}(1-\widehat{\theta})/(2n)$, respectively.)