UNIVERSITEIT VAN DIE VRYSTAAT
UNIVERSITY OF THE FREE STATE

STSB 6806

WISKUNDIGE STATISTIEK & AKTUARIËLE WETENSKAP/
MATHEMATICAL STATISTICS & ACTUARIAL SCIENCE

Test 1 — 20 April 2021

MEMORANDUM

TYD/TIME: 180 Minutes PUNTE/MARKS: 60

INSTRUCTIONS:

- Answer all questions in a single R Markdown document. Please knit to PDF or Word at the end and submit both the PDF/Word document and the .Rmd file for assessment, in that order.
- Label questions clearly, as it is done on this question paper.
- All results accurate to 4 decimal places.
- Show all derivations, formulas, code, sources and reasoning.
- Intervals should cover 95% probability unless stated otherwise.
- No communication software, no devices, and no communication capable websites may be accessed prior to submission. You may not (nor even appear to) attempt to communicate or pass information to another student.

Question 1

You are provided with a file that has the weekly market share of a company (called ACompany) for the last two years, so April 2019 to March 2021. This company has appointed you to model their market share and determine whether they have suffered or gained from lock-down.

It occurs to you that market share must be strictly between 0 and 1, so perhaps a Beta regression would be more suitable that an ordinary least squares regression.

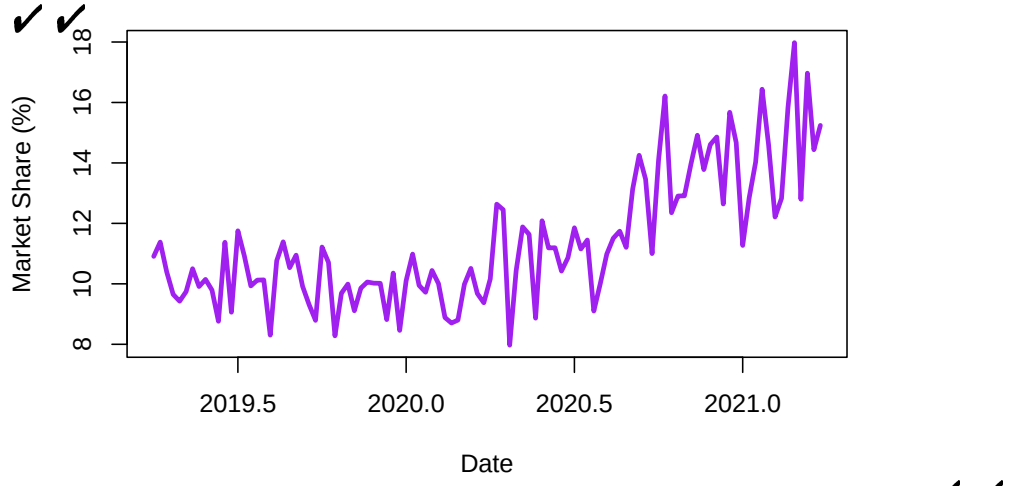
(a) Before considering any data, give the log density corresponding to $x_i \sim Beta(\alpha, \beta)$ and show that the log MDI prior is $\log p(\alpha, \beta) = (\alpha - 1)[\psi(\alpha) - \psi(\alpha + \beta)] + (\beta - 1)[\psi(\beta) - \psi(\alpha + \beta)] + \log \Gamma(\alpha + \beta) - \log \Gamma(\alpha) - \log \Gamma(\beta)$ [4]

$$\ln f(x_i) = (\alpha-1)\log x_i + (\beta-1)\log(1-x_i) + \log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta) \checkmark$$

$$E[\ln f(x_i)] = (\alpha-1)[\psi(\alpha) - \psi(\alpha+\beta)] \checkmark + (\beta-1)[\psi(\beta) - \psi(\alpha+\beta)] \checkmark$$
$$+ \log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta) \checkmark$$

(b) Load the data and draw a time series line plot with a nice scale on both axes. [4]

```
mydata <- read.csv('STSB6806Test1data2021.csv')
xts <- ts(mydata$x, start=2019.25, frequency = 52)*100
plot(xts, xlab='Date', ylab='Market Share (%)', main='', col='purple', lwd=3)
```



✔ ✔ ✔ ✔

(c) Split the data into two parts: Let $v_i = x_i,\ i = 1,\ldots,52$ be the first 52 weeks (the period before lock-down); and let $w_i = x_{i+52},\ i = 1,\ldots,52$ be the second 52 weeks (the period in lock-down). Give the mean and standard deviation of each part. Then show that the first autocorrelation of v, *i.e.* $cor(v_i, v_{i+1})$, is about 0.0313. [4]

```
v <- mydata$x[1:52]
w <- mydata$x[53:104]
cat('\n\nThe mean of $\\mathbf{v}$ is', round(mean(v),4), 'and the standard deviation
    of $\\mathbf{v}$ is', round(sd(v),4), '\n\n')
cat('The mean of $\\mathbf{w}$ is', round(mean(w),4), 'and the standard deviation of
   $\\mathbf{w}$ is', round(sd(w),4), '\n\n')
cat('The autocorrelation of $v_{i}$ and $v_{i+1}$ is', round(cor(v[1:51], v[2:52]),4)
   , '\n\n')
```

✔ ✔

The mean of **v** is 0.0995 and the standard deviation of **v** is 0.0085 ✔

The mean of **w** is 0.1276 and the standard deviation of **w** is 0.0214 ✔

The autocorrelation of v_i and v_{i+1} is 0.0313

(d) Since **v** is visually flat, the autocorrelation of **v** is very low, and the Box-Ljung

test fails to reject white noise at 6 lags, we will assume that $\mathbf{v}$ is an i.i.d. sample from a $Beta(\alpha, \beta)$ density. Fit this density using Stan, incorporating the MDI prior. Simulate at least 4000 posterior parameter vectors (Stan default). Give any summary of the model fit. [7]

```
data {
  int<lower=0> n;                    // number of observations
  vector[n] x;                         // observations
}
// The parameters accepted by the model.
parameters {
  real<lower=0> a;              // alpha
  real<lower=0> b;              // beta
}
// The model to be estimated.
model {
  x ~ beta(a, b);
  target += (a-1)*(digamma(a) - digamma(a+b)) + (b-1)*(digamma(b) - digamma(a+b)) +
      lgamma(a+b) - lgamma(a) - lgamma(b);
}
```

✔ ✔ ✔ ✔

```
library(rstan)
options(mc.cores=4)
out1 <- sampling(beta1, list(n=52, x=v))
summary(out1)$summary
```

✔ ✔ ✔

(e) Obtain posterior median estimates of each parameter, and compare them to the method of moments estimates, for $\mathbf{v}$. [2]

```
postsims1 <- extract(out1)
(alphaPostMedian <- median(postsims1$a))
(betaPostMedian <- median(postsims1$b))
ab <- mean(v)*mean(1-v)/(sd(v)^2) - 1
(alphaMOM <- mean(v)*ab)
(betaMOM <- mean(1-v)*ab)
```

✔ ✔

We observe little difference.

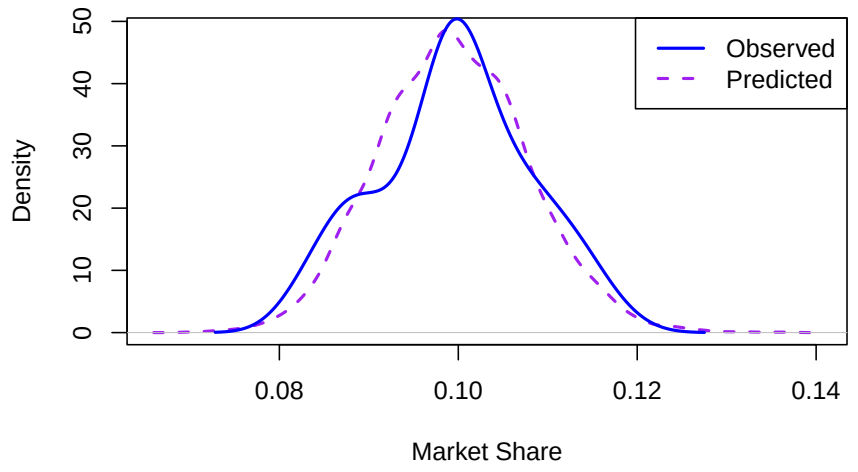
(f) Simulate at least 1 predictive value for each simulated posterior vector, to arrive at a sample from the posterior predictive density $p(v^{new}|\mathbf{v})$. Draw a kernel density plot or histogram of this density. Compare it to a kernel density plot or histogram of $\mathbf{v}$. [5]

```
nsims1 <- length(postsims1$a)
predsims1 <- rbeta(nsims1, postsims1$a, postsims1$b)
```

✔ ✔

```
plot(density(predsims1), col='purple', lwd=2, lty=2, main='', xlab='Market Share',
    ylab='Density')
lines(density(v), col='blue', lwd=2, lty=1)
legend('topright', legend=c('Observed', 'Predicted'), lwd=2, lty=1:2, col=c('blue','
    purple'))
```

✔ ✔



✔

(g) Now we move on to **w**, the period under lock-down. Model this data with a conditional Beta distribution, where the parameters change linearly over time. You may assume positive parameters, with any applicable prior (including uniform) for convenience. Hint: Do not use the MDI prior any more and see full formulation below. Give a trace plot of the simulation process for any one parameter of your choosing. [7]

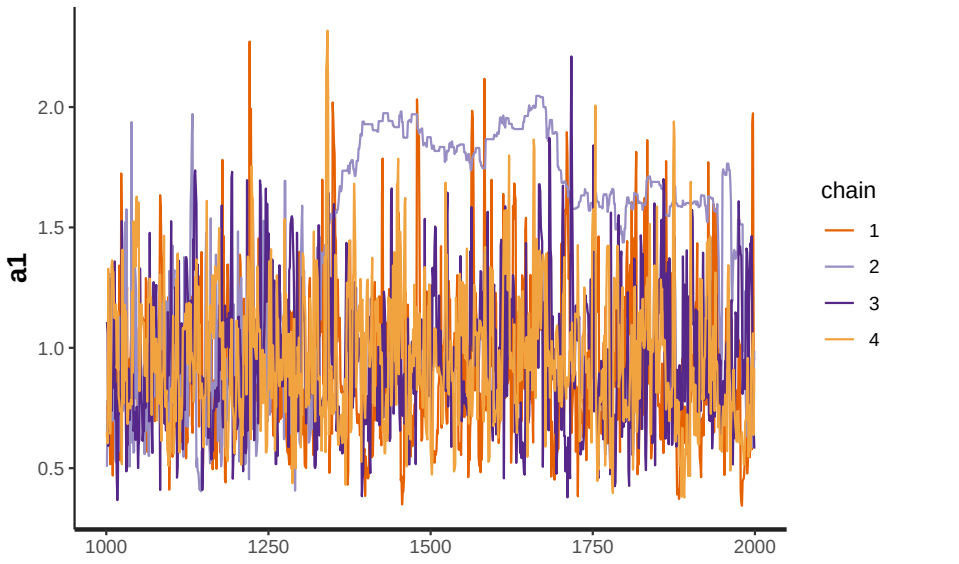
$$
\begin{aligned}
w_i &\sim Beta(\alpha_i, \beta_i) \\
\alpha_i &= a_0 + a_1 * i \\
\beta_i &= b_0 + b_1 * i \\
i &= 1, 2, \ldots, 52 \\
a_0, a_1, b_0, b_1 &> 0
\end{aligned}
$$

```
data {
  int<lower=0> n;                    // number of observations
  vector[n] x;                         // observations
  vector[n] time;
}
// The parameters accepted by the model.
parameters {
  real<lower=0> a0;              // alpha
  real<lower=0> b0;              // beta
  real<lower=0> a1;              // alpha
  real<lower=0> b1;              // beta
}
// The model to be estimated.
model {
  x ~ beta(a0 + a1*time, b0 + b1*time);
}
```

✔ ✔ ✔ ✔

```
out2 <- sampling(beta2, list(n=52, x=w, time=1:52))
traceplot(out2, 'a1')
```

✔ ✔



✔

(h) Test whether the intercept of this model for **w** is significantly different from the mean of **v**. First do this by calculating $P(\mu_v > \mu_{w0})$ where $\mu_v = \frac{\alpha}{\alpha+\beta}$ and $\mu_{w0} = \frac{a_0}{a_0+b_0}$. [5]

```
mu_v <- postsims1$a/(postsims1$a + postsims1$b)
postsims2 <- extract(out2)
mu_w0 <- postsims2$a0/(postsims2$a0 + postsims2$b0)
mean(mu_v > mu_w0)
```

✔ ✔ ✔ ✔

Around 60% probably. ✔

(i) Then calculate $P(v^{new} > w_0^{new})$ using the posterior predictive distributions of $\mathbf{v}$ and w_0. Also explain why this probability should be closer to 0.5. [4]

```
nsims2 <- length(postsims2$a0)
predsimsw0 <- rbeta(nsims2, postsims2$a0, postsims2$b0)
mean(predsims1 > predsimsw0)
```

✔ ✔

In this case we get a number more like 55%. ✔The predictive distribution accounts for additional uncertainty, and is thus less likely to indicate a difference between the distributions (more conservative). ✔

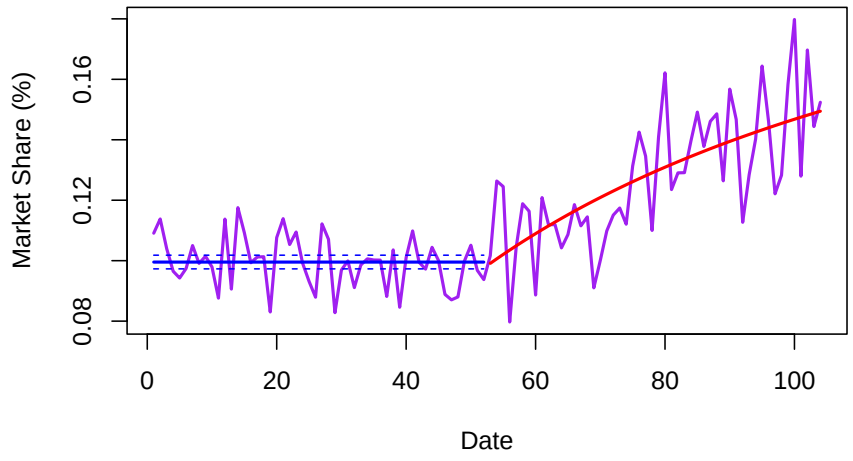
(j) Draw a plot of all the data showing the fit lines for the two models, *i.e.* a flat line going through the points for the first year at the expected value you already calculated (μ_v); and a sloped line showing the expected value fit for the second year. Also add two dotted lines showing a 95% credibility interval for μ_v for the first year. [5]

```
plot(mydata$week, mydata$x, type='l', xlab='Date', ylab='Market Share (%)', main='',
    col='purple', lwd=2)
lines(1:52, rep(mean(mu_v), 52), lwd=2, col='blue')

mu_int <- quantile(mu_v, c(0.025, 0.975))
lines(1:52, rep(mu_int[1], 52), lwd=1, lty=2, col='blue')
lines(1:52, rep(mu_int[2], 52), lwd=1, lty=2, col='blue')

a_fit <- 1:52*mean(postsims2$a1) + mean(postsims2$a0)
b_fit <- 1:52*mean(postsims2$b1) + mean(postsims2$b0)
w_fit <- a_fit/(a_fit + b_fit)
lines(53:104, w_fit, lwd=2, col='red')
```

✔ ✔ ✔ ✔



✔

(k) Fit a single joint model to all the data at once. Note that, since the first part is assumed flat, you can do this without changing the second model - you can set

$i = 0, 0, \ldots, 0, 1, 2, \ldots, 52$ and use all the data. Give a summary of the model fit. [4]

```
z <- c(rep(0,52), 1:52)
out3 <- sampling(beta2, list(n=104, x=mydata$x, time=z))
summary(out3)$summary
```

✔ ✔ ✔

Summary must make sense. ✔

(l) Give a predicted expected value and a 95% credibility interval for the predicted market share in Week 120. [4]

```
postsims3 <- extract(out3)
nsims3 <- length(postsims3$a0)
predsims3 <- rbeta(nsims3, postsims3$a0 + postsims3$a1*120, postsims3$b0 + postsims3$
    b1*120)
mean(predsims3)
quantile(predsims3, c(0.025, 0.975))
```

✔ ✔ ✔

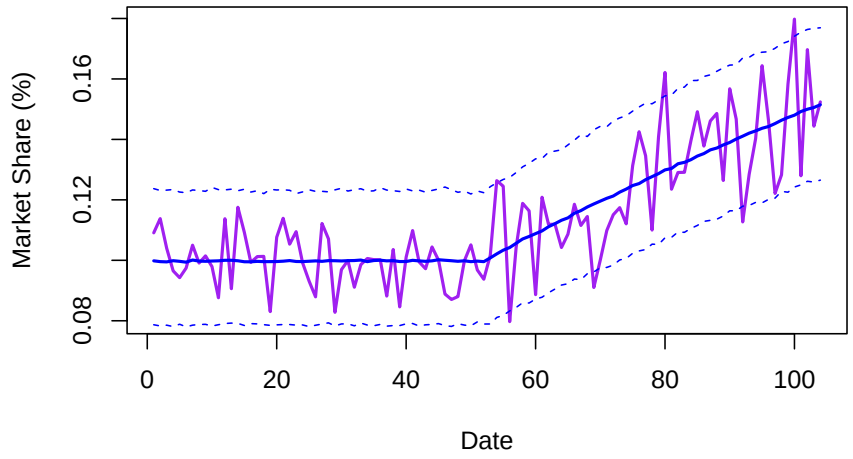
The predicted market share should be around 20% and the interval from 17% to 23.5%. ✔

(m) Draw a plot of all the data showing the predicted fit lines for the joint model. Add 95% prediction interval lines to the plot. [*If you are unable to fit the joint model then draw this plot using the separate models at a penalty of 1 mark.*] [5]

```
predmat <- sapply(z, function(i) { rbeta(nsims3, postsims3$a0 + postsims3$a1*i,
    postsims3$b0 + postsims3$b1*i) } )
means <- apply(predmat, 2, mean)
ints <- apply(predmat, 2, quantile, c(0.025, 0.975))

plot(mydata$week, mydata$x, type='l', xlab='Date', ylab='Market Share (%)', main='',
    col='purple', lwd=2)
lines(mydata$week, means, lwd=2, col='blue')
lines(mydata$week, ints[1,], lwd=1, lty=2, col='blue')
lines(mydata$week, ints[2,], lwd=1, lty=2, col='blue')
```

✔ ✔ ✔ ✔



✔

Total for Question 1: 60

Total half marks on memo = 120 vs. 120 = Double total margin points (=60).