University of Edinburgh

School of Mathematics

Bayesian Data Analysis, 2023/2024, Semester 2

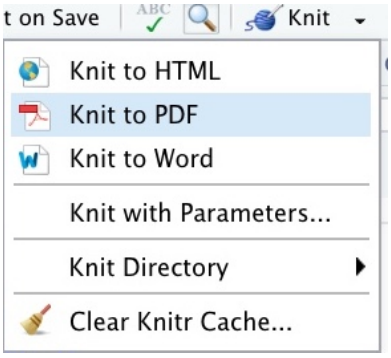
Assignment 1

IMPORTANT INFORMATION ABOUT THE ASSIGNMENT

In this paragraph, we summarize the essential information about this assignment. The format and rules for this assignment are different from your other courses, so please pay attention.

1) Deadline: The deadline for submitting your solutions to this assignment is 1 March 12:00 noon Edinburgh time.

2) Format: You will need to submit your work as 2 components: a PDF report, and your R Markdown (.Rmd) notebook (this can be in a zip file if you include additional images). There will be two separate submission systems on Learn: Gradescope for the report in PDF format, and a Learn assignment for the code in Rmd format. You need to write your solutions into this R Markdown notebook (code in R chunks and explanations in Markdown chunks), and then select Knit/Knit to PDF in RStudio to create a PDF report.



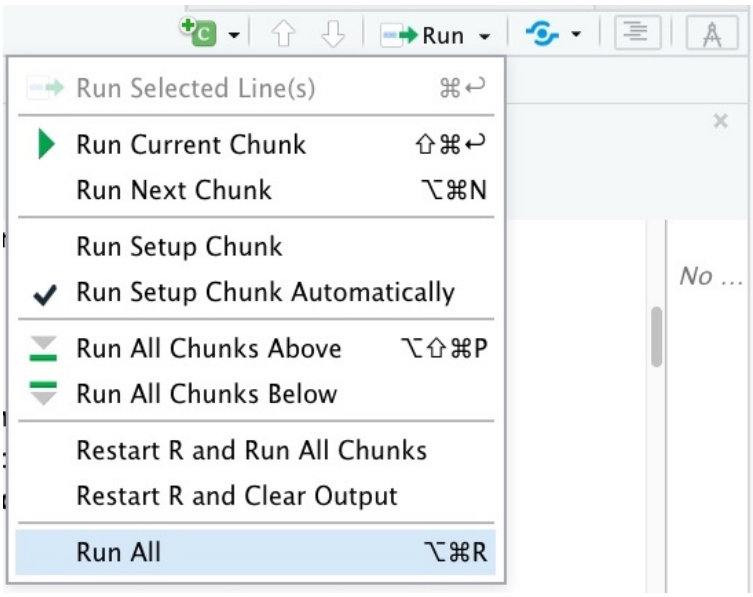
The compiled PDF needs to contain everything in this notebook, with your code sections clearly visible (not hidden), and the output of your code included. Reports without the code displayed in the PDF, or without the output of your code included in the PDF will be marked as 0, with the only feedback "Report did not meet submission requirements".

You need to upload this PDF in Gradescope submission system, and your Rmd file in the Learn assignment submission system. You will be required to tag every sub question on Gradescope.

Some key points that are different from other courses:

a) Your report needs to contain written explanation for each question that you solve, and some numbers or plots showing your results. Solutions without written explanation that clearly demonstrates that you understand what you are doing will be marked as 0 irrespectively whether the numerics are correct or not.

b) Your code has to be possible to run for all questions by the Run All in RStudio, and reproduce all of the numerics and plots in your report (up to some small randomness due to stochasticity of Monte Carlo simulations). The parts of the report that contain material that is not reproduced by the code will not be marked (i.e. the score will be 0), and the only feedback in this case will be that the results are not reproducible from the code.



c) Multiple Submissions are allowed BEFORE THE DEADLINE are allowed for both the report, and the code.
However, multiple submissions are NOT ALLOWED AFTER THE DEADLINE.
YOU WILL NOT BE ABLE TO MAKE ANY CHANGES TO YOUR SUBMISSION AFTER THE DEADLINE.
Nevertheless, if you did not submit anything before the deadline, then you can still submit your work after the deadline, but late penalties will apply. The timing of the late penalties will be determined by the time you have submitted BOTH the report, and the code (i.e. whichever was submitted later counts).

We illustrate these rules by some examples:

Alice has spent a lot of time and effort on her assignment for BDA. Unfortunately she has accidentally introduced a typo in her code in the first question, and it did not run using Run All in RStudio. - Alice will get 0 for the part of the assignments that do not run, with the only feedback "Results are not reproducible from the code".

Bob has spent a lot of time and effort on his assignment for BDA. Unfortunately he forgot to submit his code. He will get one reminder to submit his code. If he does not do it, Bob will get 0 for the whole assignment, with the only feedback "Results are not reproducible from the code, as the code was not submitted."

Charles has spent a lot of time and effort on his assignment for BDA. He has submitted both his code and report in the correct formats. However, he did not include any explanations in the report. Charles will get 0 for the whole assignment, with the only feedback "Explanation is missing."

3) Group work: This is an INDIVIDUAL ASSIGNMENT. You can talk to your classmates to clarify questions, but you have to do your work individually and cannot copy parts from other students. Students who submit work that has not been done individually will be reported for Academic Misconduct, which can lead to severe consequences. Each question will be marked by a single instructor, and submissions will be compared by advanced software tools, so we will be able to spot students who copy.

4) Piazza: During the assignments, the instructor will change Piazza to allow messaging the instructors only, i.e. students will not see each others messages and replies.

Only questions regarding clarification of the statement of the problems will be answered by

the instructors. The instructors will not give you any information related to the solution of the problems, such questions will be simply answered as "This is not about the statement of the problem so we cannot answer your question."

THE INSTRUCTORS ARE NOT GOING TO DEBUG YOUR CODE, AND YOU ARE ASSESSED ON YOUR ABILITY TO RESOLVE ANY CODING OR TECHNICAL DIFFICULTIES THAT YOU ENCOUNTER ON YOUR OWN.

5) Office hours: There will be one office hour per week (Wednesdays 16:00-17:00) during the 2 weeks for this assignment. This is in JCMB 5413. I will be happy to discuss the course/workshop materials. However, I will only answer questions about the assignment that require clarifying the statement of the problems, and will not give you any information about the solutions.

6) Late submissions and extensions: UP TO A MAXIMUM OF 3 CALENDAR DAYS EXTENSION IS ALLOWED FOR THIS ASSIGNMENT IN THE ESC SYSTEM. You need to apply before the deadline.

If you submit your solutions on Learn before the deadline, the system will not allow you to update it even if you have received an extension. There is only 1 submission allowed after the deadline.

Students who have existing Learning Adjustments in Euclid will be allowed to have the same adjustments applied to this course as well, but they need to apply for this BEFORE THE DEADLINE on the website.

https://www.ed.ac.uk/student-administration/extensions-special-circumstances

by clicking on "Access your learning adjustment". This will be approved automatically.

Students who submit their work late will have late submission penalties applied by the ESC team automatically (this means that even if you are 1 second late because of your internet connection was slow, the penalties will still apply). The penalties are 5% of the total mark deduced for every day of delay started (i.e. one minute of delay counts for 1 day). The course instructors do not have any role in setting these penalties, we will not be able to change them.

```
rm(list = ls(all = TRUE))
#Do not delete this!
#It clears all variables to ensure reproducibility
```

Ping-pong, or table tennis, is a popular sport around the world. In this assignment, we will apply Bayesian modelling to the movement of a ping-pong ball.

As explained in the paper "Optimal State Estimation of Spinning Ping-Pong Ball Using Continuous Motion Model" by Zhao et al., the physical equations describing the movement of a spinning ball in air without wind can be described as

$$\frac{dV}{dt} = k_d \|V\| V + k_m \omega \times V + g,$$

where V is the velocity of the ball (3-dimensional vector), ω is the angular velocity (3-dimensional vector), g is the local gravity acceleration (3-dimensional vector), $\times$ refers to the cross product [https://en.wikipedia.org/wiki/Cross_product]. The constants k_d and k_m are expressed as

$$k_d := -\frac{1}{2m} C_D \rho A$$
$$k_m := \frac{1}{2m} C_m \rho A r.$$

Figure 1: Ma Long. Made in China.

The meaning and values of the parameters here are shown in the following table.

TABLE I

VALUES OF THE CONSTANT PARAMETERS

Symbol	Quantity	Value
m	mass	$0.0027kg$
g	acceleration of gravity	$9.827\,m/s^2$
r	radius	$0.02\,m$
ρ	air density	$1.29kg/m^3$
A	cross sectional area	$0.001256m^2$
C_D	coefficient of air resistance	0.456
C_m	coefficient of Magnus	1.0

We observe positions and velocities at times $T_1, T_2, \ldots, T_n$, and define $\Delta_k = T_{k+1} - T_k$. The simplest way to discretize this ODE is as follows(Euler-Mayurama discretization of the original ODE, see equation (2) of "Optimal State Estimation of Spinning Ping-Pong Ball Using Continuous Motion Model"),

$$\begin{bmatrix} x(k+1) \\ y(k+1) \\ z(k+1) \\ v_x(k+1) \\ v_y(k+1) \\ v_z(k+1) \end{bmatrix} = \begin{bmatrix} x(k) \\ y(k) \\ z(k) \\ v_x(k) \\ v_y(k) \\ v_z(k) \end{bmatrix} + \begin{bmatrix} v_x(k) \\ v_y(k) \\ v_z(k) \\ -k_d\|V(k)\|v_x(k) + k_m(\omega_y v_z(k) - \omega_z v_y(k)) \\ -k_d\|V(k)\|v_y(k) + k_m(\omega_z v_x(k) - \omega_x v_z(k)) \\ -k_d\|V(k)\|v_z(k) + k_m(\omega_x v_y(k) - \omega_y v_x(k)) - g \end{bmatrix} \cdot \Delta_k,$$

where $\|V(k)\| = \sqrt{v_x(k)^2 + v_y(k)^2 + v_z(k)^2}$ is the magnitude of velocity at time T_k.

In this question, we are going to assume a similar model, but with additional Gaussian model noise, and also assume that the position and velocity observations are also subject to observation

noise. Hence, our model equations are as follows,

$$\begin{bmatrix} x(k+1) \\ y(k+1) \\ z(k+1) \\ v_x(k+1) \\ v_y(k+1) \\ v_z(k+1) \end{bmatrix} \sim N \left[\begin{bmatrix} x(k) \\ y(k) \\ z(k) \\ v_x(k) \\ v_y(k) \\ v_z(k) \end{bmatrix} + \begin{bmatrix} v_x(k) \\ v_y(k) \\ v_z(k) \\ -k_d\|V(k)\|v_x(k) + k_m(\omega_y v_z(k) - \omega_z v_y(k)) \\ -k_d\|V(k)\|v_y(k) + k_m(\omega_z v_x(k) - \omega_x v_z(k)) \\ -k_d\|V(k)\|v_z(k) + k_m(\omega_x v_y(k) - \omega_y v_x(k)) - g \end{bmatrix} \cdot \Delta_k, \Sigma \right],$$

where $\Sigma = Diag[\tau_{pos}^{-1}, \tau_{pos}^{-1}, \tau_{pos}^{-1}, \tau_{vel}^{-1}, \tau_{vel}^{-1}, \tau_{vel}^{-1}]$ is a diagonal covariance matrix depending on parameters τ_{pos} and τ_{vel}.

The observation model is the following,

$$\begin{bmatrix} x^o(k) \\ y^o(k) \\ z^o(k) \\ v_x^o(k) \\ v_y^o(k) \\ v_z^o(k) \end{bmatrix} \sim N \left[\begin{bmatrix} x(k) \\ y(k) \\ z(k) \\ v_x(k) \\ v_y(k) \\ v_z(k) \end{bmatrix}, \Sigma^o \right],$$

where $\Sigma^o = Diag[(\tau_{pos}^o)^{-1}, (\tau_{pos}^o)^{-1}, (\tau_{pos}^o)^{-1}, (\tau_{vel}^o)^{-1}, (\tau_{vel}^o)^{-1}, (\tau_{vel}^o)^{-1}]$

Q1) [10 marks]

a) [5 marks] Draw a DAG representation of this model (this can be done one a tablet, or draw it on a piece of paper and then take a picture or scan). Images can be included in R Markdown, as explained in [https://www.earthdatascience.org/courses/earth-analytics/document-your-science/add-images-to-rmarkdown-report/].

b) [5 marks] For the initial values, choose priors of the form$x(1) \sim N(0,1), y(1) \sim N(0,1), z(1) \sim N(0,1)$, $v_x(1) \sim N(0,25), v_y(1) \sim N(0,25), v_z(1) \sim N(0,25)$.Choose your own priors for τ_{pos}, τ_{vel}, τ_{pos}^o, τ_{vel}^o , ω_x, ω_y, ω_z. Explain your choices.

If you use informative priors, please cite the source of the information you used precisely (i.e. web link, or precise page number in a paper. Saying Google search said " " will not suffice).

Q2) [10 marks] In this question, we are going to fit the model of Q1) on a real dataset from [Table Tennis Ball Trajectories with Spin - Edmond (mpg.de)]. In this dataset, there are many recorded trajectories of ping-pong balls shot out by a table tennis launcher robot. We will only use one trajectory here.

First, we load the dataset, and show a 3D plot of the trajectory.

```
#If you do not have BiocManager and rhdf5 packages installed, you need to install these first.
#install.packages("BiocManager")
#BiocManager::install("rhdf5")
library(rhdf5)
#This command lists all the information in this dataset.
#Please do not include it in the knitted PDF, as it takes 20+ pages
#h5ls("MN5008_grid_data_equal_speeds.hdf5",)
n=60;
xyz.obs<-h5read("MN5008_grid_data_equal_speeds.hdf5","/originals/405/positions")[,2:(n+1)];
#Read positions of simulation number 405
xo=xyz.obs[1,];
yo=xyz.obs[2,];
zo=xyz.obs[3,];
vxvyvz.obs<-h5read("MN5008_grid_data_equal_speeds.hdf5","/originals/405/velocities")[,2:(n+1)];
```

```
#Read velocities of simulation number 405
vxo<-vxvyvz.obs[1,];
vyo=vxvyvz.obs[2,];
vzo=vxvyvz.obs[3,];

T<-h5read("MN5008_grid_data_equal_speeds.hdf5","/originals/405/time_stamps")[2:(n+1)];
#Read time points of observations

library(rgl)
rgl_init <- function(new.device = FALSE, bg = "white", width = 640) {
if( new.device | rgl.cur() == 0 ) {
  rgl.open()
  par3d(windowRect = 50 + c( 0, 0, width, width ) )
  rgl.bg(color = bg )
}
rgl.clear(type = c("shapes", "bboxdeco"))
rgl.viewpoint(theta = 15, phi = 20, zoom = 0.7)
}

rgl_init()
```

```
## Warning: 'rgl.open' is deprecated.
## Use 'open3d' instead.
## See help("Deprecated")
```

```
## Warning: 'rgl.bg' is deprecated.
## Use 'bg3d' instead.
## See help("Deprecated")
```

```
## Warning: 'rgl.clear' is deprecated.
## Use 'clear3d' instead.
## See help("Deprecated")
```

```
## Warning: 'rgl.viewpoint' is deprecated.
## Use 'view3d' instead.
## See help("Deprecated")
```

```
rgl.spheres(xo,yo,zo, r = 0.05, color = "yellow")  # Scatter plot
rgl.bbox(color = "#333377")
```

```
## Warning: 'rgl.bbox' is deprecated.
## Use 'bbox3d' instead.
## See help("Deprecated")
```

Implement the model explained in Q1) in JAGS or STAN, with the data here referring to the observations $x^o, y^o, z^o, v_x^o, v_y^o, v_z^o$.

Please treat k_m and k_d as fixed constants that can be computed based on the equations in Q1).

Evaluate the Gelman-Rubin diagnostics for model parameters τ_{pos}, τ_{vel}, τ_{pos}^o, τ_{vel}^o , ω_x, ω_y, ω_z, as well as their effective sample sizes. Choose the burn-in period, number of chains, and number of iterations such that the effective sample size is at least 1000 in all of these parameters.

Include the summary statistics from these parameters. Discuss the results.

Plot the posterior density of the angular velocity parameters (ω_x, ω_y, ω_z). Discuss the results.

Explanation: (Write your explanation here)

Q3)[10 marks] Perform posterior predictive checks on the model in Q2). Explain your choices of test functions, and interpret the results.

Do a plot of the x coordinate of the trajectory against the z coordinate, and include at least 100 of the posterior replicates on the same plot (see Line 351 in the code of Lecture 3 for a similar plot). Discuss the results.

Q4)[10 marks] In this question, we are will use the model to predict the trajectory for the next 6 time steps, and compare it to the observed values. First, we load the data including the next 6 steps (not that these observations cannot be used for prediction, only for testing).

```
n=66;
xyz.obs<-h5read("MN5008_grid_data_equal_speeds.hdf5","/originals/405/positions")[,2:(n+1)];
#Read positions of simulation number 405
xo=xyz.obs[1,];
yo=xyz.obs[2,];
zo=xyz.obs[3,];
vxvyvz.obs<-h5read("MN5008_grid_data_equal_speeds.hdf5","/originals/405/velocities")[,2:(n+1)];
#Read velocities of simulation number 405
vxo<-vxvyvz.obs[1,];
vyo=vxvyvz.obs[2,];
vzo=vxvyvz.obs[3,];

T<-h5read("MN5008_grid_data_equal_speeds.hdf5","/originals/405/time_stamps")[2:(n+1)];
```

Explain how you can implement a posterior predictive model for the position and velocity variables at the next 6 time steps, i.e. T[61], T[62],...,T[66]. Do not pass along the position and velocity observations at these time points to the model in the data (you can replace them with NA if using JAGS). Implement this in JAGS or Stan. Compute the posterior predictive mean of all position and velocity components at these new time steps, and compare them with their observed values. Compute the Euclidean distance between the observed position and the posterior predictive mean of the position variables at the next 6 time steps, and do the same for the velocities. Discuss the predictive accuracy of this Bayesian model.

Explanation: (Write your explanation here)

Q5)[10 marks] In this question, we will try to improve the model by using a different numerical discretization of the ODE

$$\frac{dX}{dt} = V, \quad \frac{dV}{dt} = k_d\|V\|V + k_m\omega \times V + g.$$

In Q1), we have used the Euler-Mayurama scheme, and added some model and observation noise.

In this question, you are expected to choose a different scheme (see e.g. Lecture 4 for some examples, or you could use the one based on the analytical solution of the ODE described in the paper "Optimal State Estimation of Spinning Ping-Pong Ball Using Continuous Motion Model"). You should still allow for model and observation noises. You can also consider different covariance matrices, such as ones that allow correlation between the model noise for x and v_x, and similarly, between the observation noise of x and v_x, etc. (such models would have additional parameters related to the amount of correlation). Describe the motivation for your scheme. Implement the new model similarly to Q2) (ESS should be at least 1000 for all model parameters), do the posterior predictive checks from Q3), and compare its predictive performance on the next 6 datapoints as in Q4). Discuss the results.

Explanation: (Write your explanation here)