Final Exam

Miodrag Bolic April 16, 2018

Instructions

- Total number of points is 135 which means that you do not need to do everything. Maximum number of point that you can get is 100.
- **Submit using Bright Space by Wednesday April 18, 2018 at 11:30pm**
- Please submit your code with the explanation. Only Python code will be accepted.
- If there are formulas or text to be written, you can write it in notebook (preferred way) or you can write on paper and then scan it and upload it.
- I saw that some students collaborated in doing assignments. However, there will be **zero tolerance** for that in the final. If I see any similarity between works (code, formulas, explanation even in one part of one question), all the students involved in that work will do 3 hour closed-book exam in my office and be sent to Academic Fraud Service to discuss their case.
- If you have questions, please send me an email at mbolic@site.uottawa.ca .

Problem 1

(25 points) (Chapter 5 from the book Measurement of Agreement: Models, Methods and Applications - available in our library and used in Lecture 8)
The data in the linked paired repeated measurement case consist of paired measurements (Y_{i1k}; Y_{i2k}) over time k = 1,...,m. Term b_{ik} represents the random effect of the common time k on the measurements. This random effect links the measurements at the time k. The model becomes:

$$Y_{i1k} = b_i + b_{i1} + b^*{}_{ik} + e_{i1k}$$
$$Y_{i2k} = \beta_0 + b_i + b_{i2} + b^*{}_{ik} + e_{i2k}$$

Let's call this model (1). The $b^*{}_{ik} \sim N(0; \sigma^*{}_b)$ and $b^*{}_{ik}$ is mutually independent of b_i, b_{ik} and e_{ik}.

Cardiac ejection fraction (in %) is measured in 12 individuals using two methods— impedance cardiography (IC) and radionuclide ventriculography (RV). The data are presented in Table below. The first column in the table is the Subject number, The second column is the result using IC and the third column is the result using RV. Columns are separated by semicolon sign ;.Both methods have an equal number of repeated measurements on an individual, but this number varies between 3 and 6. Assume that the repeated measurements are paired over time, and they are taken in the order they appear in the table.
(a) Perform an exploratory analysis of the data including scatter and Bland-Altman plots.
(b) Fit the model (1) shown above and check its adequacy. Try using PyMC3. If it does not work for you, perform fitting as we did in the assignment.
(c) Evaluate similarity and repeatability of the methods using appropriate measures.
(d) Evaluate agreement between the methods. Do the methods agree sufficiently well to be used interchangeably? If not, is there any recalibration that may bring them closer?

Table: Cardiac ejection fraction (in %).

Columns: Subject number; IC; RV

1; 6.57, 5.62, 6.9, 6.57, 6.35; 7.83, 7.42, 7.89, 7.12, 7.88
2; 4.06, 4.29, 4.26, 4.09; 6.16, 7.26, 6.71, 6.54
3; 4.71, 5.5, 5.08, 5.02, 6.01, 5.67; 4.75, 5.24, 4.86, 4.78, 6.05, 5.42
4; 4.14, 4.2, 4.61, 4.68, 5.04; 4.21, 3.61, 3.72, 3.87, 3.92
5; 3.03, 2.86, 2.77, 2.46, 2.32, 2.43; 3.13, 2.98, 2.85, 3.17, 3.09, 3.12
6; 5.9, 5.81, 5.7, 5.76; 5.92, 6.42, 5.92, 6.27
7; 5.09, 4.63, 4.61, 5.09; 7.13, 6.62, 6.58, 6.93
8; 4.72, 4.61, 4.36, 4.2, 4.36, 4.2; 4.54, 4.81, 5.11, 5.29, 5.39, 5.57
9; 3.17, 3.12, 2.96; 4.48, 4.92, 3.97
10; 4.35, 4.62, 3.16, 3.53, 3.53; 4.22, 4.65, 4.74, 4.44, 4.5
11; 7.2, 6.09, 7, 7.1, 7.4, 6.8; 6.78, 6.07, 6.52, 6.42, 6.41, 5.76
12; 4.5, 4.2, 3.8, 3.8, 4.2, 4.5; 5.06, 4.72, 4.9, 4.8, 4.9, 5.1

In [ ]:

Problem 2

(15 points)
Measurements of the speed of light are shown below.
(a) Use Kolmogorov and Anderson-Darling tests to determine normality.
(b) Using bootstrap, construct 90% confidence intervals for median and standard deviation.
(c) Perform Kernel density estimation and plot the data. What Kernel is good in your opinion and why?
(d) Model the data and measurement errors and perform Bayesian estimation of the parameters using PYMC3.

299.85 299.74 299.90 300.07 299.93 299.85 299.95 299.98 299.98 299.88 300.00 299.98 299.93
299.65 299.76 299.81 300.00 300.00 299.96 299.96 299.96 299.94 299.96 299.94 299.88 299.80
299.85 299.88 299.90 299.84 299.83 299.79 299.81 299.88 299.88 299.83 299.80 299.79 299.76
299.80 299.88 299.88 299.88 299.86 299.72 299.72 299.62 299.86 299.97 299.95 299.88 299.91
299.85 299.87 299.84 299.84 299.85 299.84 299.84 299.84 299.89 299.81 299.81 299.82 299.80
299.77 299.76 299.74 299.75 299.76 299.91 299.92 299.89 299.86 299.88 299.72 299.84 299.85
299.85 299.78 299.89 299.84 299.78 299.81 299.76 299.81 299.79 299.81 299.82 299.85 299.87
299.87 299.81 299.74 299.81 299.94 299.95 299.80 299.81 299.87

In [ ]:

Problem 3

(25 points)
The table below shows live-birth success rate in case of in-vitro fertilization. As you can see, the success rate is highly dependent on the age of the recipients. We are going to do Bayesian regression modeling in PYMC3.

(a-1) Plot data and propose your regression model and implement it in PyMC3.
(a-2) Discuss quality of fit using regression metrics that we studied.
(b) Assume change-point regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, \tau$$
$$y_i = \gamma_0 + \gamma_1 x_i + \epsilon_i, i = \tau, \ldots, n$$
$$\epsilon_i \sim N(0, \sigma^2)$$

(b-1) Propose priors (and possibly hyperpriors if you think it is better to have hierarchical model) on $\sigma^2, \beta_0, \beta_1, \gamma_0, \gamma_1$.
(b-2) Use discrete uniform prior on τ: pymc3.distributions.discrete.DiscreteUniform and write program in PyMC3 that estimates the parameters.
(c) For models in (a) and (b) analyze convergence of MCMC chains and perform modifications in case there are issues with convergences or correlation. Present metrics that we studied that are related to convergence.
(d) Compute WAIC and possibly other tests to perform model selection.
(e) Perform prediction with confidence intervals at 95% levels for the age 22.

Table
Age(x); Percentage(y)
24; 38.7
25; 38.6
26; 38.9
27; 41.4
28; 39.7
29; 41.1
30; 38.7
31; 37.6
32; 36.3
33; 36.9
34; 35.7
35; 33.8
36; 33.2
37; 30.1
38; 27.8
39; 22.8
40; 21.4
41; 15.4
42; 11.2
43; 9.2
44; 5.4
45; 3.0
46; 1.6

In [ ]:

Problem 4

(20 points)
This problem is related to very accurate measurement of mass using a balance scale. A reference weight and an additional small weight need to be put on one side to make a balance with the weight that has been measured on the other side. The goal is to measure mass of the object and to find uncertainty of the measurement.

An object of density ρ and unknown mass m is calibrated against a reference weight of density ρ_{ref} and mass m_{ref} using a balance operated in air of density ρ_{air}. The mass m_{ref} is nominally equal to 50 g, and it is known that the mass m is very close to 50 g (but it is unknown). Random variable ϵ represents the difference between ρ and ρ_{ref} and has mean of zero. Using Archimedes' principle, we find that the estimate equation is

$$m = (s + m_{ref}) \frac{1 - \frac{\rho_{air}}{\rho_{ref}}}{1 - (\frac{\rho_{air}}{\rho_{ref} + \epsilon})}$$

with s being the mass of a small weight of density ρ_{ref} added to the reference weight to give a balance.

The reference values of the parameters are:

$s = 3.8 \times 10^{-3} g,$

$\rho_{air} = 1.2 \times 10^{3} g m^{-3},$

$\rho_{ref} = 8.2 \times 10^{6} g m^{-3},$

$m_{ref} = 50 g,$

$\epsilon = 0 g m^{-3}$

and uncertainty of the parameters is defined as an error model:

$E_s \sim N(0, (0.1 \times 10^{-3})^2),$

$E_{\rho_{air}} \sim U(-0.3 \times 10^{3}, 0.3 \times 10^{3}),$

$E_{\rho_{ref}} \sim U(-0.5 \times 10^{6}, 0.5 \times 10^{6}),$

$E_{m_{ref}} \sim N(0, (5 \times 10^{-6})^2),$

$E_{\epsilon} \sim U(-1.5 \times 10^{6}, 1.5 \times 10^{6})$

This notation means that, for example, when generating Monte Carlo estimates of the variable s, you would generate random numbers from Normal distribution with the mean of $3.8 \times 10^{-3} g$ and variance of $(0.1 \times 10^{-3} g)^2$.

Find 95% confidence intervals for the mass m by performing uncertainty propagation using:
(a) Perturbation method. Please include both first and second order terms when doing Taylor series expansion.

(b) Monte Carlo method. Try to generate 1000, 10000 and 50000 random numbers and see if confidence intervals change with increasing number of samples.

Problem 5 Short unrelated questions

(35 points)

(a) Two sensors take a measurement z of a constant but unknown parameter x, in the presence of noise. Here, $z_1 = x + v_1$ with noise v_1 having distribution $N(0, \sigma_1^2)$ and $z_2 = x + v_2$ with noise v_2 having distribution $N(0, \sigma_2^2)$.
(a-1) How can one combine the two measurements to produce an optimal estimate of $\hat{x}$ of the unknown parameter x?
(a-2) What is the mean and the variance of $\hat{x}$?
(a-3) What are the mean and the variance of $\hat{x}$ in special case when $\sigma_1 = \sigma_2$?

(b) Time series
(b-1) Is IID assumption valid for time-series data and if no, why? How does that affect us when performing Bayesian analysis?
(b-2) You have a state-space model that is non-linear and the noise is Gaussian. What filter would you use to compute the estimates?
(b-3) For the problem (b-2), what parameter of your filter would you use to present the uncertainty of your estimates over time?

(c) The goal of this problem is to assess the value of a physical quantity x based on measurements that must be corrected for a background interference β. Measurement model is:

$$y = x + \beta$$

The observed values of the signal y are 3.738, 3.442, 2.994, 3.637, and 3.874. Assume based on prior knowledge and measurements of background interference that $\beta \sim Uniform(1.126, 1.329)$. y is modeled as normal distribution with standard deviation σ_y. The prior for σ_y is $\sigma_y \sim Uniform(0, c)$. Select c as well as other parameters so that they make sense for this problem based on the data that is given. Estimate the parameters of interest using Bayesian method.

(d) Uncertainty propagation is performed using correlated input data.
(d-1) How would that affect computation of uncertainty of the output using perturbation method?
(d-2) How would that affect computation of uncertainty of the output using Monte Carlo method?

(e) List MCMC methods that we studied and analyze their advantages and disadvantages

(f) Consider the model $y = x_1 \times x_2 + x_3 \times x_4$ where $x_i \sim N(0, \sigma_i)$ for $i = 1, \ldots, 4$. Find Sobol indexes S_i and S_{ij}

Problem 6

(15 points)
a) Load data from the file GP_Data.csv. The first raw represents values of x and the second raw values of y. Plot the data.

b) Fit the data using Gausssian Processes with Squared Exponential Kernel. Plot the fitting with confidence intervals.

c) Fit data using the Kernel of your choice.Plot the fitting with confidence intervals.

d) Predict the values of the output for x=52, x=60 and x=70 with confidence intervals.

In [ ]: