Exam - Reproducible data treatment with R

2019/12/03

Download the Exam.zip file and unzip it in a folder called "`Rexam`".

Open the `exam.Rmd` file and rename it as `NAME_FirstName.Rmd`. Put your name in the header of the file.

You will write all your text and code in this Rmd file, and you will email (colin.bousige@univ-lyon1.fr) to me the file at the end of the exam. **The file should compile (knit) with no error**. Unless specifically specified, all graphs can be done either using base graphics or `ggplot2`.

Internet access and research is *authorized*.

Sharing your answers through email, Facebook or any other mean *is not*.

Exercise 1 (4 points)

1. Print the 6 first lines of the R-built-in data.frame `trees`
2. Print only the column names
3. What is the dimension of `trees`?
4. Plot the trees height and volume as a function of their girth in two different graphs. Make sure the axis labels are clear
5. In each graph, add a red dashed line corresponding to the relevant correlation that you observe (average value, linear correlation...)
6. Explain your choice and write the corresponding values (average value and standard deviation, or slope, intercept and corresponding errors). Round the values to 2 decimals.

Exercise 2 (6 points)

1. Print the 3 first lines of the R-built-in data.frame `USArrests`. This data set contains statistics about violent crime rates by US state. The numbers are given per 100 000 inhabitants, except for `UrbanPop` which is a percentage.
2. What is the average murder rate in the whole country?
3. What is the state with the highest assault rate?
4. Create a subset of `USArrests` gathering the data for states with an urban population above (including) 80%.
5. How many states does that correspond to?
6. Within these states, what is the state with the smallest rape rate?
7. Print this subset ordered by decreasing urban population.
8. Print this subset ordered by decreasing urban population and increasing murder rate.
9. Plot an histogram of the percentage of urban population with a binning of 5%. Add a vertical red line marking the average value. Make sure the x axis shows the [0,100] range.
10. Is there a correlation between the percentage of urban population and the various violent crime rates? argument your answer with plots.

Exercise 3 (10 points)

In high-pressure experiments, the pressure in the Diamond Anvil Cell (DAC) is calibrated through the measure of the Raman shift of a tiny ruby crystal placed in the pressure transmitting medium next to the measured sample.

1. Write a function returning the pressure P as a function of the ruby Raman shift position ω and the excitation laser wavelength λ_l:

$$P(\omega, \lambda_l) = \frac{A}{B}\left[\left(\frac{\lambda}{\lambda_0}\right)^B - 1\right](GPa)$$

   where $A = 1876$ and $B = 10.71$, λ is the the measured wavelength of the ruby R_1 line (the most energetic one) and $\lambda_0 = 694.24$ nm is the zero-pressure value at 298 K [1]. The relationship between the wavenumber ν in cm^{-1} and the wavelength λ in nm is given by $\nu(\text{cm}^{-1}) = \frac{10^7}{\lambda(\text{nm})}$, and the Raman shift $\omega = \Delta\nu = \nu_l - \nu = \frac{10^7}{\lambda_l} - \frac{10^7}{\lambda}$ (cm^{-1}).

2. Write a function returning a normalized Lorentzian as a function of its center x_0 and its full width at half maximum Γ:

$$L(x) = \frac{\Gamma}{2\pi}\frac{1}{\frac{\Gamma^2}{4} + (x - x_0)^2}$$

3. Store the list of files containing `ruby` in their name in the `Data/` folder into a variable `flist`. Print its length.

4. Plot with points the first file in `flist`. Find the position of its maximum and store it in `xmax`. Guess roughly the parameters needed to fit the experimental data by `y0+A1*L(x,x1,FW1)+A2*L(x,x2,FW2)`, and add a blue line on the plot to represent this function.

5. Using `nls()`, fit the first spectrum in `flist` by `y0+A1*L(x,x1,FW1)+A2*L(x,x2,FW2)` and using the starting parameters you defined before. Plot the experimental data again and add the fitted spectrum as a red line.

6. Based on the above procedure, for each file in `flist` (so, use a `for` loop), fit the Raman spectrum by the sum of two Lorentzian functions, and store the fitting parameters into a data.frame called `ruby_fit` also containing the names of the corresponding files. Attention: the initial guesses for amplitudes and widths can be constant, but the peaks positions should evolve for each spectrum. The difference between the two peaks is always roughly 30 cm^{-1}, and the largest peak is always the most energetic one. Check that your fits are correct by printing the experimental data and the fitted result at each iteration (add the name of the file as the plot title).

7. Add a column in `ruby_fit` corresponding to the estimated pressure rounded to 1 decimal. The excitation wavelength in this experiment was 532 nm. Print the resulting `ruby_fit` table using `knitr::kable(ruby_fit)`

8. Store all file names containing "RBM" into a variable `fRBM`. Load all the corresponding spectra into a single `data.frame` called `spec` with 3 columns: Raman shift ω, Intensity, Pressure. Of course, the indexes in the file names between the ruby and RBM files match. In the Intensity column, store the intensity normalized to [0,1].

9. Using `ggplot2`, plot with points the stacked normalized RBM band spectra vertically shifted by P, with a color for each spectrum corresponding to the pressure. Make the plot interactive.

References

[1] Chijioke *et al.* 'The ruby pressure standard to 150 GPa'. *J Appl Phys* **98,** 114905 (2005). DOI: 10.1063/1.2135877