Advanced Regression Assignment 1

30 April, 2020

Instructions

1. Please hand in both an R markdown file and a compiled document version of your R markdown file on Vula. Your compiled document should not contain any code. If you need to describe an algorithm this should be done in text or in the form of pseudo-code.
2. Your report should contain enough information, description of methods, interpretation of results and conclusions to explain clearly what you have done. It should be clear what you have done without having to refer to any code. This means that you need to specify all settings used, even if these are default settings. A good guideline is to give enough detail to allow another person to replicate what you have done exactly, even if they want to replicate this using a different software package.
3. Please attach a plagiarism statement to your hand-in. Your code and write-up need to be entirely your own work, even though you are allowed to discuss the work with others. Please submit your document to Turnitin.
4. The assignment is due: 27 May 2020.

Question 1: P-Splines

The data for this question is on a set of 892 females under 50 years collected in three villages in West Africa (triceps.csv). The aim here is to understand how body fat (measured roughly as triceps skinfold thickness) changes with age (use log of skinfold thickness).

Start with a cubic B-spline basis and 20 evenly spaced (interior) knots. Adding a difference penalty (squared difference) to adjacent coefficients will create P-splines. Fit a P-spline to the triceps data. Calculate the GCV score for a range of values of the smoothing parameter λ. Choose a suitable value for the smoothing parameter, and illustrate your results. Outline your methodology.

Question 2: Thin-plate Splines

Refer to the paper by Simon Wood on thin-plate regression splines: Wood, S.N., 2003. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B*, 65(1), pp.95-114.

Section 3.1 and Appendix A: Replicate the comparison in section 3.1.:

1. Simulate from the model by simulating 100 points on the unit square, evaluate at the function and add noise: $e \sim N(0, 0.1^2)$.
2. Choose 16 evenly spaced knots, construct a thin-plate spline basis, and fit this to the data (not penalized).
3. Construct a full thin-plate spline basis. Reduce its rank using an eigenvalue deconstruction (Appendix A) to a 16 rank basis. Fit this to the data.
4. Plot the fitted surfaces, and compare the two methods. Describe the methodology used.

$d = 2$, dimension of surface.

m: order of differentiation in the wiggliness penalty. Choose $m = 2$.

T contains basis functions for constant and linear terms: The $M = \binom{m+d-1}{d}$ functions are linearly independent polynomials spanning the space of polynomials with degree less than m (pg. 97).

E contains radial basis functions associated with knots

Use the following radial basis functions: $r(u) = u^2 \log(u)$, where u is Euclidean distance between observation and knot.

In the file thin-plate-low-rank.Rmd are some guidelines for following the steps in the Wood appendix. There is also code for predicting the surface after having fitted the reduced rank thin-plate spline. My surface is the wrong way around, … just so that there is something left for you to figure out. (I don't currently know the answer, but let me know if you find it please). The predicted points are the right way around, so it must be something in the surface predictions.

I have added references, so you can check where the equations come from.

One thing I might not have done exactly as in Wood is that I have used SVD instead of an eigen-decomposition. Seems to work.

Question 3: GAMs and MARS

The data here are on the quality of red wine. The measurements relate to red variants of the Portuguese "Vinho Verde" wine. For details see the reference Cortez et al., 2009. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). The data were obtained from Kaggle. The data are in file `winequality-red.csv`.

The data can be analysed via classification or regression. The classes are ordered and not balanced (e.g. there are more normal wines than excellent or poor ones).

For this project investigate which properties determine the quality of wine. Choose a cutoff and treat quality as a binary variable, 1 for good, 0 for not so good. Use both GAMs and MARS to find a model which will allow you to predict the quality of wine. Clearly describe all settings / specifications you have used. You can use R's GAM functions.

Illustrate your results. Compare GAMs and MARS with respect to fitted curves, and predictive accuracy. Interpret your results and give conclusions.

Question 4: Wavelets

For this question you should fit a regression model which includes both spline terms and wavelet terms. The idea comes from work by Anestis Antoniadis (who gave a seminar earlier this year on exactly this topic, although his talk was mostly about penalties and thresholding for those penalties) and coauthors (see references).

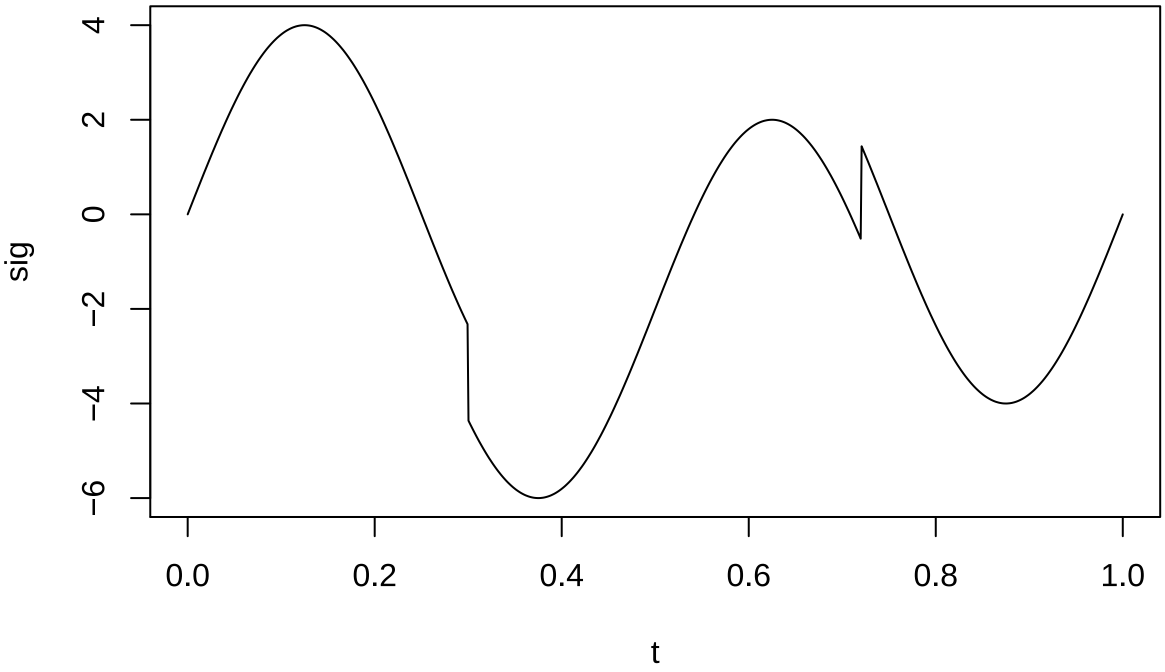
The function in the following figure is called a HeaviSine function. Most of the curve is smooth and can be well modelled using splines, however splines will have a hard time in recovering the two sharp edges.

```r
## Heavisine
## matlab code MakeSignal.m Wavelab

t <- seq(0, 1, length = 1000)

sig = 4*sin(4*pi*t)
sig = sig - sign(t - .3) - sign(.72 - t)

plot(sig ~ t, type = "l")
```



In reality you would only obtain noisy observations on the signal. If we add noise, using a signal to noise ratio of 3, our data would look as follows:

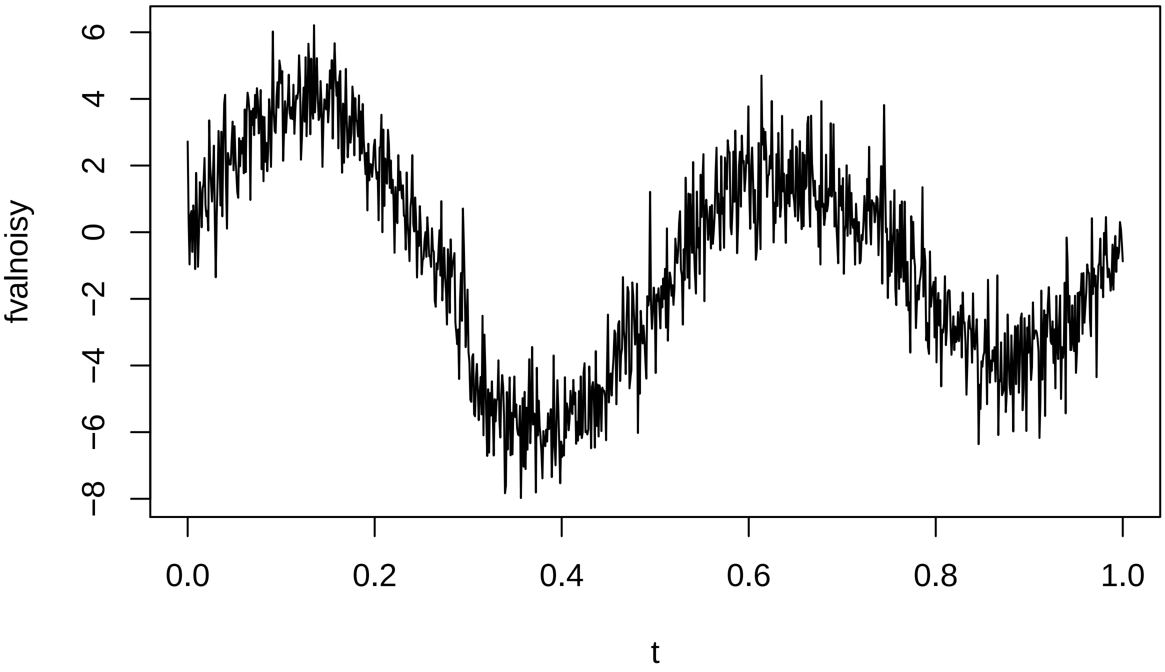
```
rsnr = 3

sdnoise = sd(sig)/rsnr

epsilon= rnorm (length(sig)) * sdnoise

#signal + noise
fvalnoisy = sig + epsilon

plot(fvalnoisy ~ t, type = "l")
```



The signal to noise ratio was calculated as $\frac{sd(f)}{\sigma}$ where σ is the standard error of Gaussian white noise.

Fit a regression model with spline and wavelet terms to recover the signal. Choose one type of wavelet (Antoniadis used Symmlet 8 and Coiflet 3 wavelets). You can either choose to only penalize / threshold the the wavelet coefficients or both wavelet and spline coefficients. Try doing this as part of a penalized least squares approach.

Show plots which decompose the observations into smooth terms, wavelet terms and noise.

Include methodology and results in your report.

Question 5: Functional Data Analysis

The Southern Oscillation Index (SOI) is one of many large-scale weather indicators. There is a correlation between this index and rainfall in the summer rainfall area of South Africa, not so much with the winter rainfall are.

Here you should use *functional regression* to model total annual rainfall in terms of daily SOI (scalar response = annual rainfall, functional predictor daily SOI). Monthly rainfall for Bloemfontein Airport is available in the file `bloem_monthly_rain.csv`. Daily SOI data is available in the file file `DailySOI1887-1989Base.txt` (SOI data from Queensland Government, rainfall data from NOAA.)

It probably makes sense to sum rainfall for a summer rainfall season, instead of for one year.

For this question you should not use any specialized R functions for functional data analysis, but you can use R's regression functions and create spline bases using R's functions.

Illustrate results, carefully outline your methodology, and interpret your results.

References

1. Antoniadis A, Bigot J, Sapatinas T. 2001. Wavelet Estimators in Nonparametric Regression: A Comparative Simulation Study. J Stat Soft 6(6).
2. Antoniadis A. 2007. Wavelet methods in statistics: some recent developments and their applications. Statist Surv. 1(0):16–55.

3. P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems 47:547-553.