6404 Take-home Exam 3: Wavelets, Categorical Data Analysis, Nonparametric Regression

Peter Williams, pwilliams60@gatech.edu

Date: 2018-11-27

Contents

1. Wavelets (50%)

Locate a one-dimensional data set that has sharp-changes like those presented in the recent lectures for applying the following wavelet procedures. It is best that the data size is larger than 512, and is in 2-factorial, e.g., 210 = 1024. Note that you need to locate proper R-package/codes to perform the tasks below.

i) Select two families of wavelets for completing the two tasks ii) and iii) below, and make comparisons for the results impacted from distinct wavelet families.

ii) Show a multi-resolution plot of mother and father discrete wavelet-coefficients (DWTs), and make comments about their values.

iii) Apply two thresholding/shrinkage methods to reduce number of non-zero DWT coefficients. Apply the IDWT to reconstruct the original data signal by using the thresholded DWTs. Comment on the quality of the reconstructions from plotting the original and reconstructed signals (by overlaying them in one figure). Compare the two methods about their data-reduction ratios and the MSEs.

iv) Artificially alter the data in one "local-region" and one large-size "global region" for creating 3 distinct "fault-class" data sets. See lecture presentation about details of this task. Apply the best thresholding method (and the best wavelet-family) to model the data from all FOUR classes (one from the original data and the other 3 fault-class data). Use the multi-resolution plots of thresholded-DWTs to see how they are different in these FOUR classes of data signals.

v) Discuss the possibility and steps for developing a rigorous decision-making procedure to detect faulty-signals against the original data, and distinguish classes of faulty signals with reduced-size data presented by thresholded DWTs.

2. Categorical Data Analysis (25%)

Chapter 9 of the textbook on categorical data analysis includes 6 sections with various problems/data and methods. Locate three sets of problems/data matching three distinct methods taught in lectures. Apply proper statistical software (preferred to be in R-codes) to analyze the data. Provide in-depth comments about the findings in your statistical analyses.

1. Chi-square, goodness of fit

To discuss the Chi-square goodness of fit procedure, a dataset published by the *National Oceanic and Atmospheric Administration* was located that tracks the number of hurricanes that have made landfall on the continental United States by decade over the last ~ 60 years. Links to the actual source data is linked below. To get a sense of what the dataset looks like a brief preview is shown below, where the expected count of hurricanes is computed as the grand mean of observed landfalls across the entire dataset:

Table 1: NOAA: Continental United States: Hurricane Impacts/Landfalls

Decade	Count of Impacts	Expected Impacts
1951-1960	18	15.667
1961-1970	15	15.667
1971-1980	12	15.667
1981-1990	16	15.667

Decade	Count of Impacts	Expected Impacts
1991-2000	14	15.667
2000-2010	19	15.667

Source: Continental United States Hurricane Impacts/Landfalls by decade as reported by the NOAA http://www.aoml.noaa.gov/hrd/hurdat/All_U.S._Hurricanes.html

Given the overall national interest in climate change, and its impact on weather patterns, it is of interest to many researchers if the frequency of hurricane landfalls in recent years in increasing. Given the dataset shown above, we can test this hypothesis in a crude way utilizing differences in observed vs. expected frequency of hurricane landfall. Where n_i refers to observed counts across $i = 1, ..., k$ decades measured, and $N = \sum_{=1}^{k} n_i$, is the total number of observed landfalls. We can then set up our test:

$H_0 : F_X(x) = F_0(x)$
$H_\alpha : F_X(x) \neq F_0(x)$

And then take our test statistic T to be, $T = \sum_i^r \frac{(n_i - np_i)^2}{np_i}$, which is approximately distributed $T \sim \chi^2_{r-1}$. Jumping into actual calculations, we have $T = \frac{(18-np_1)^2}{np_1} + \frac{(15-np_2)^2}{np_2} + ... + \frac{(19-np_6)^2}{np_6} = 2.128$, where $np_i = 15\frac{2}{3}$, for $i = 1, ..., 6$.

Under our H_0, our p-value is 0.831, not sufficient evidence to reject our H_0 and conclude that hurricane landfalls are different across decades, but perhaps a contentious talking point in conversation. Obviously the scope and depth of the analysis here isn't sufficient to add meaningfully to the overall discussion about climate change.

2. Contingency tables, testing for homogeneity and independence

To demonstrate usage of contigency tables and associated tests using *R*, an article published by medical researchers that followed $N = 6272$, Swedish men for 30 years to see whether there was any association between the amount of fish in their diet and prostate cancer was located. According to authors of the research, the original study actually used pairs of twins. This approach allowed researchers to share their findings with more confidence.

In the table below, we show frequency counts among research respondents across their fish diet type, and whether or not they contracted prostate cancer across the 30 year study:

Table 2: Occurence of Prostate Cancer, By Amount of Fish Consumption, N = 6272

	Large	Moderate	Small	Never
No	507	2769	2420	110
Yes	42	209	201	14

Source: Lancet, June 2001, "Fatty Fish Consumption and Risk of Prostate Cancer"

The table itself does not communicate readily any differences in cancer rates by diet type. To evaluate further how diet type may impact the rate of cancer, we utilize the chi-square test of independence, which is readily available via the function *chisq.est* in base *R*. Given a contingency table of consisting of m levels of a factor, along with an additional factor with k levels of observations, we can construct a matrix $N = (n_{ij}), i = 1, ..., m,\ j = 1, ..., k$, where each entry in N is an observed frequency, or count. The main gist of the chi-square test, is to study observed frequencies in N, n_{ij}, and their potential differences from the expected (under the hypothesis of independence) frequencies, denoted: $\hat{n_{ij}} = \frac{n_{i\cdot} \cdot n_{j\cdot}}{n_{\cdot\cdot}}$, where $n_{i\cdot} = \sum_{j=1}^{k} n_{ij}$, and $n_{j\cdot} = \sum_{i=1}^{m} n_{ij}$, and compute our test statistic $T = \sum_{i=1}^{m} \sum_{j=1}^{k} \frac{(n_{ij} - \hat{n_{ij}})^2}{\hat{n_{ij}}}$, where $T \sim \chi^2$. The results of the

test for our fish diet dataset are summarized in the table below:

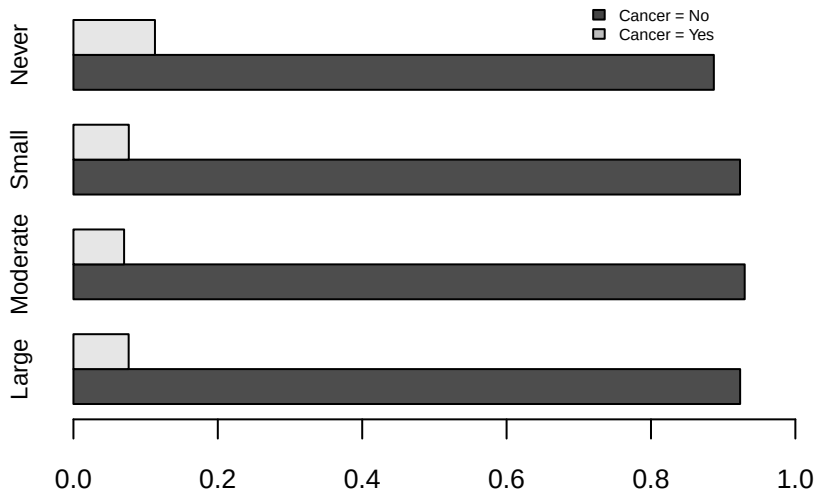
Table 3: Chi-square Test for Independence Test, R (base), function: "chisq.test (df = 3)"

Result	Value
Test Statistic (Chi)	3.677
P-value	0.298

As shown, our resulting P-value 0.298 does not provide evidence for us to conclude that the observed frequency of prostate cancer by fish diet type, is unexpected by itself. While this finding may be interesting, this test alone isn't sufficient to rule out diet's role in respondent's contraction of prostate, as there are potentially many other factors which can contribute to prostate cancer in addition to diet alone.

3. Fisher exact test

Taking our analyis of prostate cancer, and fish diet further, visual inspection of the marginal proportion of prostate cancer by fish diet classification shows that those who never consumed fish in their diet, may have had slightly higher rates of prostate cancer, which was identified and estimated in the *Lancet* paper referenced above. Here is a basic barplot to visualize difference in rate of cancer by diet:



Since, as shown in the text, the Fisher exact test is based on the null hypothesis that *two* factors, each with *two* factor levels, are independent, conditional on fixing marginal frequencies for both factors. However, the fish diet dataset described above is 2×4, therefore for the purposes of this question, we subset our data to just those with 'Never', and 'Large' fish diets for comparison.

As shown below, relying on the only the *fisher.test* function to compute *Fisher's Exact Test* in *R (base)* to detect any differences in counts across diet categories. The test shown below, relies on the odds ratio between the occurence of cancer between groups ('Never' and 'Large'):

Table 4: Fisher's Exact Test: Fish Diet Comparison (Confidence = 0.95)

Result	Value
Odds Ratio	1.535

Result	Value
Upper Confidence Level	0.747
Lower Confidence Level	2.989
P-Value	0.207

Again, our resulting P-value from *Fisher's Exact Test* does not provide evidence for us to conclude that the observed frequency of prostate cancer by fish diet type, is unexpected. As highlighted before, there are potentially many other factors which can contribute to prostate cancer in addition to diet alone. The research referenced above in *Lancet* takes a deeper dive into this particular dataset and concludes that diet, in fact, did play a role describing respondents' rate of contracting prostate cancer, however, relying on more sophisticated statistical techniques.

3. Nonparametric Regression (25%)

Locate a data set suitable for both kernel and spline regressions. The data should include at least 3 x-variables. It is okay to focus on additive-models discussed in lectures.

To consider procedures for both kernel and spline regression, we utilize the well-known automobile MPG dataset that is available in the UCI Machine Learning library. Link here: https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data. The dataset was used in the 1983 American Statistical Association Exposition, and is also well referenced in the following paper:

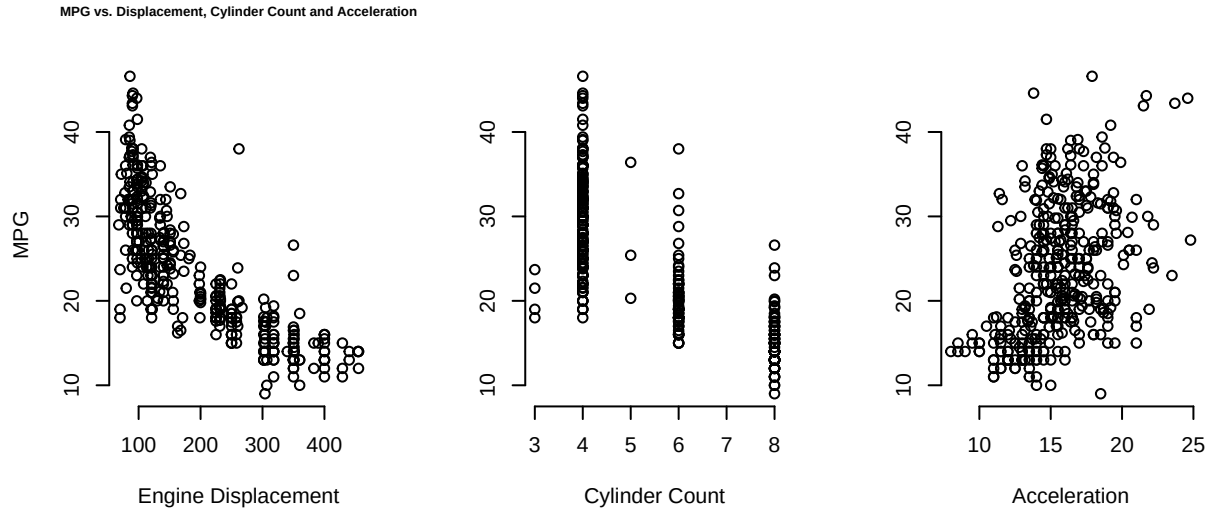
Quinlan,R. (1993). Combining Instance-Based and Model-Based Learning. In Proceedings on the Tenth International Conference of Machine Learning, 236-243, University of Massachusetts, Amherst.

As highlighted by Quinlan in the paper above, this dataset is concerned with city-cycle fuel consumption in miles per gallon, to be predicted in terms of a number of attributes of each car. Each car in the dataset ($N = 392$) was manufactured sometime in between 1970 and 1982. For convenience purposes we have removed observations with missing values. Based on the exercise specification we will rely on a car's *Mileage Per Gallon* (mpg), as our outcome of interest, and utilize 3 co-variates, namely the car's *Cylinder* count (cyl), *Engine Displacement* (disp), and *Acceleration* (accel). A preview of the data is provided below:

Table 5: Preview of Automobile MPG Dataset (Published 1993)

MPG	Cylinder Count	Displacement	Acceleration	Model
18	8	307	12.0	chevrolet chevelle malibu
15	8	350	11.5	buick skylark 320
18	8	318	11.0	plymouth satellite
16	8	304	12.0	amc rebel sst
17	8	302	10.5	ford torino
15	8	429	10.0	ford galaxie 500

One aspect of this dataset that makes it interestingly suite to kernel and spline regression, is that visually, there is evidence of non-linear relationship between the outcome of interest (*MPG*) and *Engine Displacement*. The relationship between *MPG*, *Cylinder* count and *Acceleration* is less clear, and shown below:



1) Go through one-variable-at-a-time kernel- and spline-regression fits to the data for all 3 x-variables. Compare the fitting results using the leave-one-out cross-validation procedure.

2) Select 2 sets of x-locations, e.g., (1st set: x1 = 3, x2 = 5, x3 = 2, where 3, 5, 2 are values within x-data range). These 2 sets of x-locations should be from a location close to the center of x-data-range and another location closer to the edge. Make predictions of Y at these x-data. Compare the predictions from Kernel and spline methods, and also comment on the impact from x-data-locations.

3) Construct the 90% Bias-Corrected Bootstrap CIs for the predictions at the selected 2-sets of x-locations. Show details of the bias-correction process. Compare the CIs at two x-locations, and also from two nonparametric regression methods.

Code Appendix