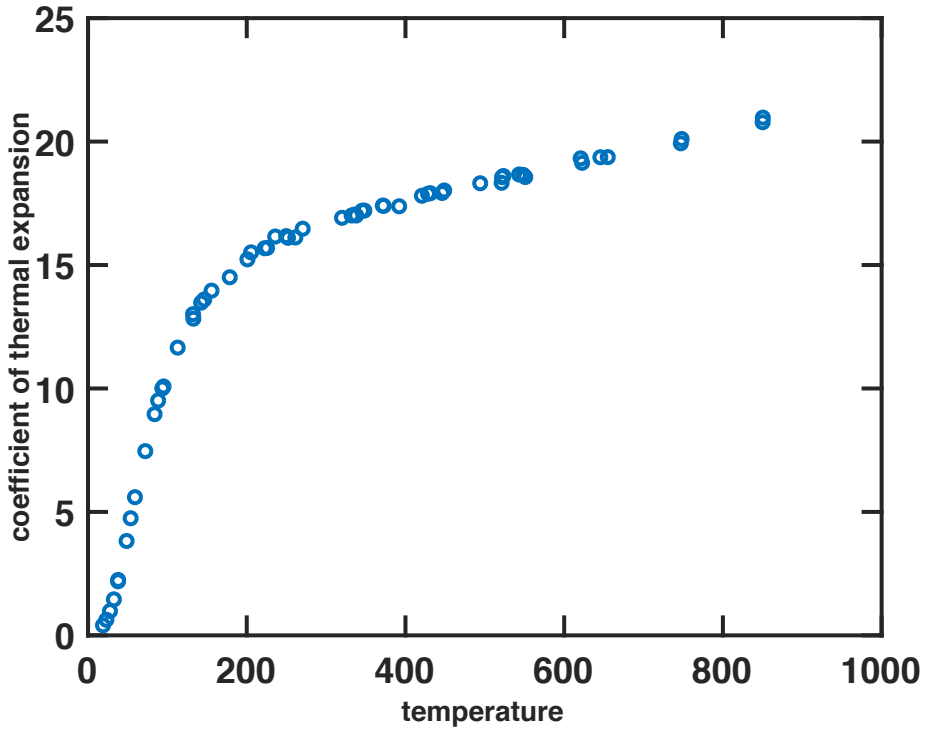
ISyE 6416: Computational Statistics
Homework 5
(100 points total.)

This homework is due on **April 6, 2020**.

- Please write your team member's name if you collaborate.

1. **Nonlinear regression using spline.** (40 points.)

   The coefficient of thermal expansion y changes with temperature x. An experiment to relate y to x was done. Temperature was measured in degrees Kelvin. (The Kelvin temperature is the Celcius temperature plus 273.15). The raw data file is copper-new.txt



   (a) Perform linear regression on the data. Report the fitted model and the fitting error.

   (b) Perform nonlinear regression with spline function (i.e., using all data points). Use GCV for find λ. Report the fitting error.

   (c) Predict the coefficient at 400 degree Kelvin.

2. **PCA for face recognition.** (30 points.)

(a) Perform data analysis on the Yale face dataset (in Canvas) for subject 14. Plot the mean face and the first 6 eigenfaces for subject 14.

(b) Now use subject14.test.gif to perform face recognition using the following procedure.

For doing face recognition through PCA we proceed as follows. Given the test image, we project it using the first component to obtain the coefficient vector, and compare $|z^\intercal u_{1,i}|$, $i = 1, 2$, where $u_{1,i}$ is the first dominant component for ith person. Are we able to recognize the person correctly using the first principle component?

Remark: you have to perform downsampling of the image by a factor of 4 to turn them into a lower resolution image before we do anything. See the example MATLAB code.

3. **Recommender systems.** (30 points)

Take our movie recommender systems data (which is obtained from a course survey in previous year), and perform recommendation using the following methods. Hint, when using features, expanding the categorical features using "one-hot" keying.

(a) User-based collaborative filter to recommend 5 movies to each user. Try several similarity metrics, which are defined in the forms of

$$\text{sim}(u, v) = e^{-(d(u,v))^2},$$

where

i. (1) $d(u, v) = \ell_2$ distance of the features, i.e., if x_u is the feature vector of the first movie, and x_v is the feature vector of the second movie, then $d(u, v) = \|x_u - x_v\|_2$.

ii. (2) $d(u, v) = \ell_1$ distance of the features, i.e., $d(u, v) = \|x_u - x_v\|_1$;

iii. (3) $d(u, v) =$Hamming distance of the features, i.e., $d(u, v) = \|x_u - x_v\|_0$;

(b) Item-based recommendation to recommend 5 movies to each user. try using the above three metrics respectively.

(c) Using matrix completion algorithm based on soft-impute (R package `https://cran.r-project.org/web/packages/softImpute/softImpute.pdf`)[1] to fill out missing entries to recommend 5 movies to each user.

Report all of the three findings in an excel spread sheet (with 3 tabs for each of methods), and each row is in the format of (name, recommended movie 1, score of the recommended movie 1, ..., recommended movie 5, score of the recommended movie 5).

[1] More reference here `https://web.stanford.edu/~hastie/swData/softImpute/vignette.html`