Hw1

Problem1

1. The cumulative distribution function of Z is given by:
   $F_z(z) = \int_{x1-x2\le z} f_1(x_1)f_2(x_2)dx_1dx_2$
   $= \int_{-\infty}^{\infty}\int_{-\infty}^{x1-z} f_1(x_1)f_2(x1-z)dx_1dx_2$
   $= \int_{-\infty}^{\infty} f_1(x1)F_2(x1-z)dx_1$
   Take the derivative on both side,
   $f(z) = \int_{-\infty}^{\infty} f_1(x_1)f_2(x1-z)dx_1$.

2. For X_1 and X_2 in [0,1], $X_1 - X_2 = z$ is in [-1,1];
   If z is negative, $f(z) = \int_{-\infty}^{\infty} f_1(x_1)f_2(x_1-z)dx1 = f(z) = \int 1*(1+z)dx1$;
   If z is positive, $f(z) = \int_{-\infty}^{\infty} f_1(x_1)f_2(x_1-z)dx1 = f(z) = \int 1*(1-z)dx1$;
   So $f(z) = (1-|z|)$, -1<=z<=1

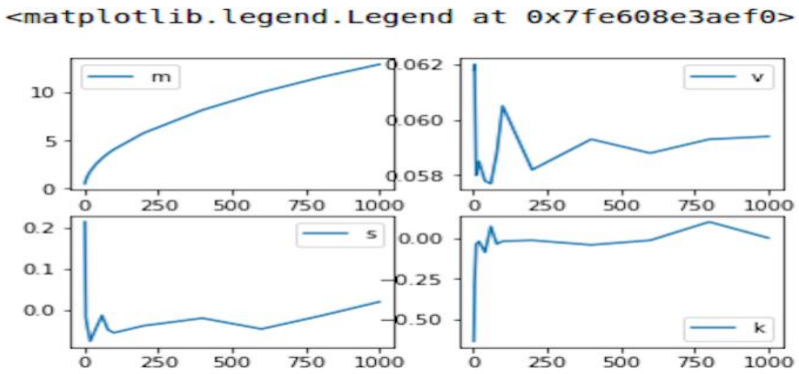
3. Because $Z = X_1 - X_2$, so the Euclidean distance between X_1 and X_2 are $|Z|$.
   $E(|Z|) = \int|z|*f(z)dz = \int_{-1}^{1}|z|f(z)dz = 2*\int_0^1|z|(1-|z|)\,dz$
   $= 2\int_0^1(z^2-z^3)dz = 2*(\frac{1}{2}-\frac{1}{3}) = \frac{1}{3}$

4. $t = z^2$, then $E(|z^2|) = \int z^2*f(z)dz = \int_{-1}^{1} z^2 f(z)dz = 2*\int_0^1 z^2\,(1-|z|)\,dz$
   $= 2\int_0^1(z^3-z^4)dz = 2*(\frac{1}{3}-\frac{1}{4}) = \frac{1}{6}$

5. The simulation can be seen in Jupyter Notebook.

6. N = [2,5,10,20,40,60,80,100, 200, 400, 600, 800, 1000]
   mean= [0.52,0.88,1.27,1.80,2.57,3.15,3.65,4.07,5.76,8.16,10,11.55,12.9]
   variance =
   [0.0618,0.0620,0.0580,0.0585,0.0578,0.0577,0.0588,0.0605,0.0582,0.0593,0.0588,0.0593,0.0594]
   skew= [0.2133, -0.0176, -0.0393, -0.0745, -0.0425, -0.0124, -0.0470, -0.0542, -0.0373, -0.0189, -0.0450, -0.0134, 0.0206]
   kurtosis = [-0.6382, -0.2795, -0.0358, -0.0181, -0.0856, 0.0745, -0.0345, -0.0167,-0.0113, -0.0398, -0.0113, 0.1033, 0.0034]

7. This is the graph for mean, variance, skewness and kurtosis given N.
   (0, 0): mean; (0,1): variance; (1,0): skewness, (1,1): kurtosis



Problem2

1. The data is divided into a training set and a testing set. Model coefficients are generated based on the training set before applied to the testing set.

2. I extended the code by putting all analysis in a loop which contains all stock in a list. Estimations of all stocks are presented below. Graphs are shown in ipynb file.

   ----------Resulting for company: HON----------
   M2.1 Lorek and Willinger (1996) via scikit-learn
   training data set has 49 observations from 1981-06-30 to 1993-06-30
   model coefficients: [[0.1168978 0.04711579 -0.22233131 -0.10432182 -0.03632673
   0.37489571
     0.7420571 -1.59811807]]
   model intercept: 1.9562
   scikit-learn in-sample RMSE = 0.7695
   scikit-learn out-of-sample RMSE = 1.9869

   M2.1 Lorek and Willinger (1996) via scikit-learn
   training data set has 96 observations from 1981-06-30 to 2005-03-31
   model coefficients: [[0.01035651 0.0218203 -0.08661981 -0.09730454 -0.03424837
   0.28276715
     0.53685904 -0.12622538]]
   model intercept: 0.3473
   scikit-learn in-sample RMSE = 0.6833
   scikit-learn out-of-sample RMSE = 0.4152

----------Resulting for company: JPM----------
M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 49 observations from 1981-06-30 to 1993-06-30
model coefficients: [[-9.90093714e-01 -1.35061907e-01 1.01939008e-02 3.75517044e-03
  -2.60902411e-14 1.11477419e+00 -2.61722355e+00 -5.51013348e+00]]
model intercept: 2.8484
scikit-learn in-sample RMSE = 3.6425
scikit-learn out-of-sample RMSE = 1.3862

M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 96 observations from 1981-06-30 to 2005-03-31
model coefficients: [[-1.07079420e+00 -8.94565020e-02 1.08688339e-02 5.19364123e-05
  -3.99680289e-15 1.22038447e+00 -2.23308760e+00 -3.51943938e+00]]
model intercept: 2.0063
scikit-learn in-sample RMSE = 2.6929
scikit-learn out-of-sample RMSE = 5.7926

----------Resulting for company: XOM----------
M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 49 observations from 1981-06-30 to 1993-06-30
model coefficients: [[0.08657576 0.10658962 -0.17050735 0.07912351 0.03319268 0.13185495
   0.19726108 0.50991298]]
model intercept: -0.0633
scikit-learn in-sample RMSE = 0.3021
scikit-learn out-of-sample RMSE = 0.3462

M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 96 observations from 1981-06-30 to 2005-03-31
model coefficients: [[0.17576334 0.03821836 -0.14764905 0.08455256 0.02379087 0.05482103
   0.27666968 0.88790183]]
model intercept: -0.1026
scikit-learn in-sample RMSE = 0.2929
scikit-learn out-of-sample RMSE = 0.6184

----------Resulting for company: HAL----------
M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 49 observations from 1981-06-30 to 1993-06-30
model coefficients: [[0.10694884 0.35430395 -0.47920709 0.00123896 0.18882622 0.49750713
  -0.20725946 -0.72584845]]

model intercept: 1.1844
scikit-learn in-sample RMSE = 0.7445
scikit-learn out-of-sample RMSE = 1.2497

M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 96 observations from 1981-06-30 to 2005-03-31
model coefficients: [[0.12130866 0.26549499 -0.13539567 0.02154886 0.17207587
0.01972469
  -0.00368412 -0.30949834]]
model intercept: 0.1164
scikit-learn in-sample RMSE = 0.6425
scikit-learn out-of-sample RMSE = 0.8226

----------Resulting for company: HOV----------
M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 43 observations from 1985-02-28 to 1995-04-30
model coefficients: [[0.24821493 0.3563095   0.00208424 0.01485516 0.08286148 -
0.0197722
  -0.05028437 0.397699]]
model intercept: -0.2380
scikit-learn in-sample RMSE = 0.2273
scikit-learn out-of-sample RMSE = 0.6509

M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 86 observations from 1985-02-28 to 2006-01-31
model coefficients: [[0.41983379 0.46576563 0.00720489 0.00054185 -0.01754442
0.00111906
  -0.00889981 0.052974]]
model intercept: -0.0686
scikit-learn in-sample RMSE = 0.3359
scikit-learn out-of-sample RMSE = 1.7702

----------Resulting for company: INTC----------
M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 49 observations from 1981-06-30 to 1993-06-30
model coefficients: [[0.5170455 -0.0440158 -0.06638112 -0.23286886 0.00122009
0.19499043
   0.44616838 0.04196248]]
model intercept: 0.0889
scikit-learn in-sample RMSE = 0.2674
scikit-learn out-of-sample RMSE = 0.3770

M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 96 observations from 1981-06-30 to 2005-03-31

model coefficients: [[-0.11385532 0.03954753 -0.14860962 -0.0303687   0.00158788
0.57158118
  -0.26810383 0.01653162]]
model intercept: -0.0496
scikit-learn in-sample RMSE = 0.2840
scikit-learn out-of-sample RMSE = 0.1582

----------Resulting for company: IBM----------
M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 49 observations from 1981-06-30 to 1993-06-30
model coefficients: [[-0.32932787 0.88976148 0.14217504 -0.36477359 -0.15631829
1.11080542
  -0.64244446 -1.33882841]]
model intercept: -1.8381
scikit-learn in-sample RMSE = 2.0937
scikit-learn out-of-sample RMSE = 6.4315

M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 96 observations from 1981-06-30 to 2005-03-31
model coefficients: [[0.14749682 0.01689593 0.01102821 -0.05397064 0.05666986
0.39335386
  -0.54750355 0.89388823]]
model intercept: 1.2613
scikit-learn in-sample RMSE = 2.0139
scikit-learn out-of-sample RMSE = 1.7405

----------Resulting for company: L----------
M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 49 observations from 1981-06-30 to 1993-06-30
model coefficients: [[-4.76539617e-02 7.07892270e-02 -3.53955305e-02 -7.05831241e-
04
   7.16383917e-02 -3.34142584e-02 1.28446890e-01 1.91492520e+00]]
model intercept: 2.1477
scikit-learn in-sample RMSE = 2.0114
scikit-learn out-of-sample RMSE = 3.1735

M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 96 observations from 1981-06-30 to 2005-03-31
model coefficients: [[0.09718525 -0.01371855 -0.22111952 -0.00069162 0.03255112 -
0.00170762
   0.65501718 0.49284154]]
model intercept: 1.1079
scikit-learn in-sample RMSE = 2.4482
scikit-learn out-of-sample RMSE = 1.2353

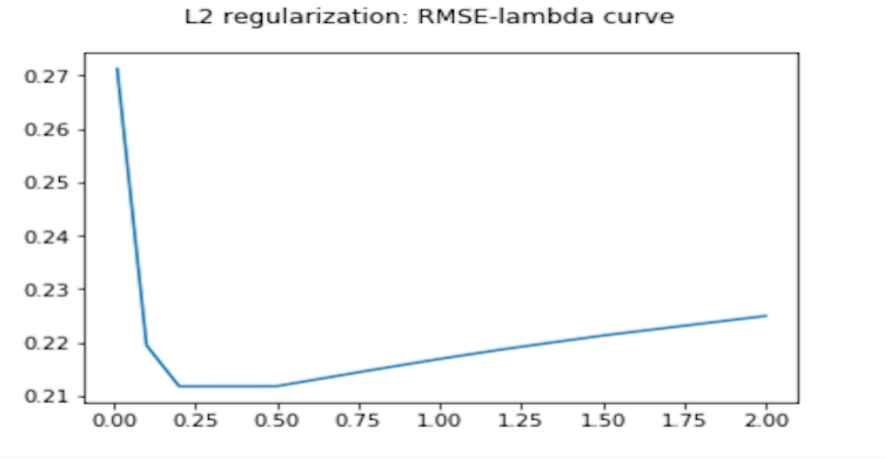
----------Resulting for company: MDT----------
M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 49 observations from 1981-04-30 to 1993-04-30
model coefficients: [[0.14463692 0.0209364   0.05722302 0.03447081 -0.01455055
0.03467077
  0.00095313 -0.0655434]]
model intercept: 0.2219
scikit-learn in-sample RMSE = 0.1482
scikit-learn out-of-sample RMSE = 0.1458

M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 96 observations from 1981-04-30 to 2005-01-31
model coefficients: [[0.26582252 0.12288498 0.06015176 0.02724939 -0.01312361
0.0361634
 -0.0286583 -0.01044202]]
model intercept: 0.1016
scikit-learn in-sample RMSE = 0.1410
scikit-learn out-of-sample RMSE = 0.2315

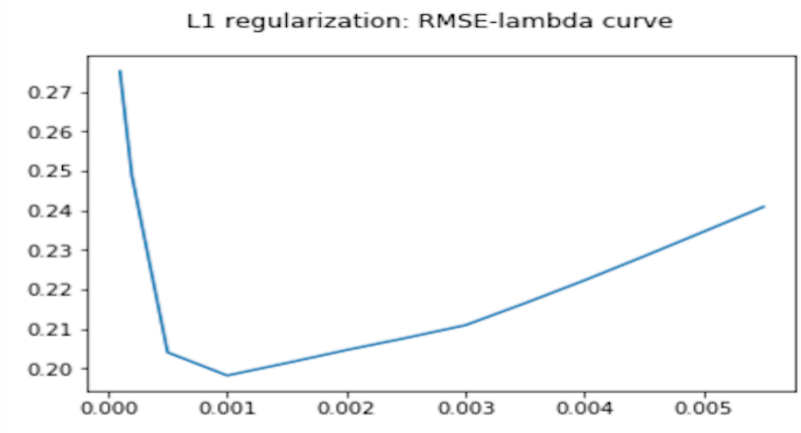
----------Resulting for company: MSFT----------
M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 41 observations from 1987-06-30 to 1997-06-30
model coefficients: [[1.03647099 0.09328008 -0.24930418 0.17806657 -0.18108243 -
0.67058393
  0.95034594 -1.49187651]]
model intercept: 0.7729
scikit-learn in-sample RMSE = 0.1203
scikit-learn out-of-sample RMSE = 0.2919

M2.1 Lorek and Willinger (1996) via scikit-learn
training data set has 80 observations from 1987-06-30 to 2007-03-31
model coefficients: [[3.98571236e-01 2.85126910e-01 2.63514962e-04 3.13690192e-
02
 -1.16365064e-01 -2.15001048e-01 5.37484362e-01 5.30088650e-01]]
model intercept: -0.1125
scikit-learn in-sample RMSE = 0.1346
scikit-learn out-of-sample RMSE = 0.3162

3. For L2 regularization, I used lambda from list [0.01, 0.1, 0.2, 0.5, 0.8, 1.0, 1.2, 1.5, 2.0] in the Ridge regression. I get the MSE corresponding to all lambda: [0.271, 0.219, 0.211, 0.211, 0.214, 0.216, 0.218, 0.221, 0.224]. The plot is shown below:



4. For L1 regularization, I used lambda from list [0.0001, 0.0002, 0.0005, 0.0010, 0.0015, 0.0020, 0.0025, 0.0030, 0.0035, 0.0040, 0.0045, 0.005, 0.0055] in the Lasso regression. I get the MSE corresponding to all lambda which is the list [0.275, 0.248, 0.204, 0.198, 0.201, 0.204, 0.207, 0.210, 0.216, 0.222, 0.228, 0.234, 0.240] and the plot is shown below:



5. When I tried different values of lambda, I found that when lambda = 0.0035 or lambda = 0.0055, the first and second training model results have only 4 non-vanishing coefficients respectively.