Group Project

1. *The project is due 20 November (Monday of Reading Week).*
2. *You have two options for the group project. Both options are equally good.*
   (a) ***Option A:*** *You can choose a topic on your own related to what we covered in class.*
      - *A possible topic: "Does LLM help with the credit decision in the LendingClub case? How?"*

   (b) ***Option B:*** *Alternatively, you can finish a project following the instructions of this document. (The instructions will hopefully also give you a reference point regarding the project scope even if you pick Option A.)*
3. *Please submit the following documents in a single zip file through Canvas (only one submission per group):*
   - *Your typed answers in a single write-up file;*
   - *Your source codes (if any);*
   - ***Additional documents****:*
      - *If you choose Option A, please record a 10 min presentation video explaining your project and the main results. (You can use Zoom to do this.)*
      - *If you choose Option B, please submit your final recommendation of the portfolio (for the data challenge) stored in a CSV file.*
         * *This CSV file should only have one column with 43 entries. The header of this column should be "G#", where "#" is your group number on Canvas.*
         * *Please name this file "Recommendation_G#.csv", where "#" is your group number.*

   Please name your zip file in the format of "GP_G#.zip", where "#" is your group number on Canvas. For example, if you are submitting for Group 1, please name your zip file as "GP_G1.zip".
4. *References to resources not in the textbook or class handouts should be explicitly mentioned in the write-ups and source codes.*
5. *Your submission will be graded based primarily on your reasoning and correctness. Presentation (e.g., clarity of writing, visualization of data, etc.) will also be considered.*
6. *The group project has 50 points in total.*

Good Luck!!!

The file *gp_data_1986_to_2015.csv* contains the following information (Credit: Kenneth R. French Data Library):

- *Mkt-RF*: The excess return on the market portfolio, i.e., the value-weight return of all applicable firms incorporated in the US and listed on the NYSE, AMEX, or NASDAQ;
- *RF*: The riskless return rate;
- *The next 43 columns*: The monthly return of 43 industry capital-weighted portfolios. The descriptions of those industries are provided in the file *43_industry_description.txt*.

All returns are measured in the unit of percentages on a monthly basis. These data series all start in January of 1986 and run through the end of 2015. Therefore, there are 30 years in total.

1 Data Exploration and Analysis (15 pts)

First, we will reinforce the tools we have learned in the class and construct diversified portfolios among the 43 industries.

1.1 Pre-processing

1. Calculate the excess return of the 43 industry portfolios. (Note that the excess return of the market portfolio is already provided.) Subsequently, we will work with the excess returns directly and treat the riskless rate as zero. Also, you may also treat each industry portfolio as an individual asset.

1.2 Basic Portfolio Construction and In-sample analysis

1. Suppose you are asked to allocate your investment among the 43 industries. Use the 1986-2015 historical data, construct the following three portfolios:
   - The equally weighted portfolio (EWP);
   - The tangency portfolio (TAN);
   - The global minimum variance portfolio (GMV).
2. Calculate the realized (in-sample) expected returns, standard deviations, Sharpe ratios, and betas of every 43 industry portfolio, as well as the market portfolio (MKT), EWP, TAN, and GMV.

Deliverable.

1. In a 3 by 4 table, summarize the three in-sample performance metrics (i.e., expected return, standard deviation, and Sharpe ratio) of the four portfolios (MKT, EWP, TAN, and GMV).

2. Plot the following two graphs related to in-sample analysis:

   (a) *The σ vs. $E[r]$ diagram.* This diagram includes all of the 43 industry portfolios, the special portfolios (i.e., EWP, TAN, and GMV) constructed from the 30-year data, the market portfolio (MKT), and the in-sample efficient frontier.

   (b) *The β vs. $E[r]$ diagram.* This diagram includes the same set of portfolios (the 43 industry portfolios treated as risky assets, MKT, EWP, TAN, and GMV) and a straight line that represents the relationship between β and $E[r]$ implied by CAPM (also known as the security market line).

1.3 Robust Portfolio Construction and Out-of-sample Analysis

First, we will consider the following robust estimation of parameters:

- *Estimating Beta.* We will use the following shrinkage estimator of beta:

$$\beta_{shrink} = 0.5\bar{\beta} + 0.5\hat{\beta}.$$

  In the expression above, $\bar{\beta} = \frac{1}{43}\sum_{i=1}^{43}\hat{\beta}_i$ is the grand average of beta across industries, $\hat{\beta}$ is the beta vector directly estimated from the sample covariance with MKT, and 0.5 is the shrinkage constant that has been determined for you.

- *Estimating Expected Return.* Since estimating the expected return of an asset is a notoriously difficult task, we will impose two layers of robustness in the estimation:

  – First, we will use the CAPM-based expected return. That is the expected return implied by CAPM that only depends on the value of beta.

  – Second, we will use the robust estimation of beta itself; please see above.

  For the easiness of reference, let us call this estimation of return μ^{CAPM}.

- *Estimating covariance matrix.* We will use the following shrinkage estimator of covariance matrix:

$$V_{shrink} = 0.3V^{CC} + 0.7\hat{V}.$$

  In the expression above, V^{CC} is the *constant correlation matrix* estimated from data, $\hat{V}$ is the sample covariance matrix, and 0.3 is the shrinkage constant that has been determined for you.

Second, with the robust estimation of parameters, we may construct the following robust version of portfolios:

- *TAN-robust*: The tangency portfolio using V_{shrink} and μ^{CAPM} as problem inputs;
- *GMV-robust*: The global minimum variance portfolio using V_{shrink} as the problem input.

In other words, the (only) difference between TAN and TAN-robust is the input parameters (i.e., expected return and covariance matrix), and we do not disturb the mathematical optimization problem itself. A similar thing can be said between GMV and GMV-robust.

Finally, we can compare the six different portfolio constructions (MKT, EWP, TAN, TAN-robust, GMV, and GMV-robust) in an out-of-sample study.

In this section, let us perform a simple version of the out-of-sample analysis. That is, we break the data into two components:

- The training data correspond to the first 25 years, i.e., years 1986-2010.
- The test data correspond to the last 5 years, i.e., years 2011-2015.

In the out-of-sample analysis, we construct the portfolios from the training data but evaluate them on the test data.

Deliverable.

1. *The out-of-sample σ vs. $E[r]$ diagram.* This diagram includes all of the 43 industry portfolios (treated as risky assets), MKT, and the special portfolios constructed from **training data** (EWP, TAN, GMV, TAN-robust, and GMV-robust). All of them are evaluated on the **test data**. Also, please include the "true" and "realized" EF curves.

2. Compare the out-of-sample performances of MKT, EWP, TAN, TAN-robust, GMV, and GMV-robust using the definitions and methods introduced above. Report your results in a 3 by 6 table. (Recall that we have 3 performance metrics and 6 portfolios to compare.)

3. Contrast the table for out-of-sample performance with that for in-sample performance. What are your findings and insights? Also, are there any limitations to the current evaluation methodology?

2 The Data Challenge (35 pts)

Imagine you are an investor at the beginning of the Year 2016 and would like to know the answer to the following single question:

- How to allocate my money among the 43 industries with an investment horizon of 5 years?

With the provided historical data, try to come up with a portfolio that you recommend.

2.1 Rules of the Data Challenge

1. Your final recommendation is a vector of weights for the 43 industries.
2. Your weights among the 43 industry portfolios should sum up to one.
3. You could use the toolboxes in this class, or you can develop your own methods. (If you use external sources, please explicitly mention them.)
4. For simplicity, the weights are static: you cannot re-balance your portfolio within the next 5 years.
5. You cannot use any information you are not supposed to know at the beginning of the Year 2016 (for example, you should not Google the stock prices in 2018). Use your best judgment on this rule!

2.2 Deliverable

Please submit your final recommendation. Please also summarize your thought process and reasoning for your recommendations. Report the intermediate steps/calculations/visualizations, if any.

2.3 Evaluation

1. Your recommendation will be evaluated on the 2016-2020 data (which was not made available to you). The relevant performance metrics are the average return, standard deviation, and Sharpe ratio.
2. The goal of this data challenge is to give you a flavor of portfolio management in practice. Your grades in this section will be mainly based on how thorough your thought process is and your understanding of the subject. The 2016-2020 performance will only take a small portion of grades, but the top-performing team will get bonus points.