

## MATH-GA.2711 Machine Learning & Computational Statistics

### Homework 4

Profs. Ivailo Dimov

Due: Feb 22, 2026

### Instruction

This homework is to be done individually. No collaboration and/or code sharing permitted.

### Methodology and Deliverables

1. Performance analysis of hedge fund returns using linear regression.

(a) Download and import the following data into your Jupyter notebook:

- Monthly data of the Fama-French 5 Factor model from Ken French's data library. The Fama-French 5 Factor model is described [here](#).
- Monthly returns from Yahoo Finance for the AQR Equity Market Neutral Fund Class I (Ticker: QMNIX). Make sure you get the series that is adjusted for both dividends and splits (i.e. the series marked "Adj Close\*\*" on Yahoo Finance). Note that Yahoo Finance will provide you with prices. Hence you will have to compute returns yourself.

Use the longest time period you can find where all data series are available.

- (b) Create scatter plots and compute summary statistics of all data series (mean, std, skew, kurtosis, max, min, correlation matrix). Discuss your conclusions and any noteworthy results after visualizing and exploring the data. (Extra credit: Perform tests for stationarity of all data series.)
- (c) Regress QMNIX on the Fama-French factors and discuss the regression results, including  $t$ - and  $F$ -tests and  $R^2$ . Is the intercept statistically different from 0?
- (d) Demonstrate graphically whether the classical linear regression assumptions are satisfied or not in (c).
- (e) For students in the Data Science in Quantitative Finance class (extra credit for everyone else): Provide a financial/economic interpretation of your results from (c).

- (f) Extra credit: Download and import the monthly data of the 10 Industry Portfolios model from Ken French's data library. The 10 Industry Portfolios model is described here. Then perform the analysis (c)-(e) using this data.
- (g) Extra credit: Download and import the *daily* data of all the data series. Make sure you use the same time period as above. You will find them on Ken French's website and Yahoo Finance. Then perform the same analysis as above. Are the classical linear regression assumptions satisfied for the regressions with daily data? Does your analysis using daily data provide different results compared to the analysis based on monthly data? Why, or why not?

## 2. Bias-variance decomposition.

- (a) Derive the bias-variance tradeoff for MSE error

$$\mathbb{E}[(\hat{\beta} - \beta)^2] = \text{var}(\hat{\beta}) + (\hat{\beta}^* - \beta)^2.$$

- (b) Simulate the model

$$y = \beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2),$$

with fixed  $(\beta, \sigma^2)$  and  $B \geq 10,000$  Monte Carlo repetitions. Consider the shrinkage estimator

$$\hat{\beta}_\lambda := \frac{1}{1 + \lambda} y, \quad \lambda \geq 0.$$

For a grid of  $\lambda$  values (e.g.  $\lambda \in \{0, 0.1, 0.2, \dots, 5\}$ ), compute empirically (1) the bias (2) the variance and (3) the MSE error when averaging over the Monte Carlo draws. Verify numerically the bias-variance tradeoff above for each  $\lambda$  and plot the MSE versus  $\lambda$

- (c) For the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

with orthogonal regressors, show that

$$\sum_{j=1}^p \mathbb{E}[(\tilde{\beta}_j - \beta_j)^2] = p\sigma^2 \left( \frac{1}{1 + \lambda} \right)^2 + \left( \frac{\lambda}{1 + \lambda} \right)^2 \sum_{j=1}^p \beta_j^2.$$

where the shrinkage estimator  $\tilde{\beta}_j$  is defined as

$$\tilde{\beta}_j := \frac{1}{1 + \lambda} \hat{\beta}_{\text{OLS},j}. \quad j = 1, \dots, p, \quad \lambda \geq 0.$$

Derive the optimal shrinkage value

$$\lambda^* = \frac{p\sigma^2}{\sum_{j=1}^p \beta_j^2}.$$

- (d) Generate an orthogonal design matrix  $\mathbf{X} \in \mathbb{R}^{n \times p}$  (with  $n \gg p = 4$ ) (how would you do that without the QR decomposition). Check that  $\mathbf{X}'\mathbf{X} = \mathbf{I}_p$ . Choose a fixed  $\boldsymbol{\beta} \in \mathbb{R}^p$  and  $\sigma^2 > 0$ , simulate

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n),$$

and compute  $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ . For a grid of  $\lambda$  values as in part (b), compute  $\tilde{\boldsymbol{\beta}}(\lambda) = \frac{1}{1+\lambda} \hat{\boldsymbol{\beta}}_{\text{OLS}}$  and estimate the empirical MSE

$$\widehat{\text{MSE}}(\lambda) := \frac{1}{B} \sum_{b=1}^B \left\| \tilde{\boldsymbol{\beta}}^{(b)}(\lambda) - \boldsymbol{\beta} \right\|_2^2$$

over  $B$  Monte Carlo repetitions. Compare by empirically minimizing  $\lambda$  to the theoretical  $\lambda^*$  above. Summarize results in a Pandas `DataFrame` with columns `lambda`, `mse`, and identify the minimizing `lambda`. Plot `mse` as a function of `lambda` and mark  $\lambda^*$  on the plot.