Big Data Asset Pricing: Exercises

© Theis Ingerslev Jensen and Lasse Heje Pedersen*

This version: April 27, 2023

These exercises are part of a course on Big Data Asset Pricing. The exercises train students in working with big asset pricing data in light of economic theory, including beta-dollar neutral portfolio construction, constructing value factors, factor replication analysis and multiple testing adjustments, high-dimensional return prediction, and research proposals.

See course website for more information:
https://www.lhpedersen.com/big-data-asset-pricing

*Jensen is at Copenhagen Business School. Pedersen is at AQR Capital Management, Copenhagen Business School, and CEPR; www.lhpedersen.com. AQR Capital Management is a global investment management firm, which may or may not apply similar investment techniques or methods of analysis as described herein. The views expressed here are those of the authors and not necessarily those of AQR.

How to Do These Exercises

These exercises are an essential part of the course on Big Data Asset Pricing. Students taking the class for credit must hand in their solutions of all exercises. The class is graded on a pass/fail basis and satisfactory completion of each of these exercises is needed to pass the class.

Each student must individually produce a satisfactory solution to each exercise and submit it by the deadline (by uploading it as an assignment in Canvas). Each student must also be able to explain the solution and present it.

Students are allowed to discuss the exercises and solution methods, but students are not allowed to copy each other. Students must should disclose any other material used, for example if code has been copied from public sources (adapting public code is perfectly fine, but should be disclosed). More broadly, students must behave according to the Code of Professional Conduct and Ethics.[1]

The solution of each exercise must consist of a self-contained PDF file with the following elements:

- **A cover page with**
  - full name and affiliation (university and department)
  - the following signed statement: "I certify with my signature that I have solved the exercise according to the Code of Professional Conduct and Ethics. For example, I have not plagiarized others, but, instead, solved the exercise myself (possibly with allowed collaboration with other students), and I have referenced my sources appropriately."
- **The body of the solution**
  - A written description of the solution with discussion, tables, and figures.
  - Think of your written description as a short research paper. Importantly, your solution should include your approach as well as the final results. For example, if asked to create a factor and plot its cumulative return, you need to show the plot *and* describe how you created the factor and how you interpret the result. In other words, the exercise is also meant to train you in how to present research in figures, tables, and writing (so you are ready to write a great paper).
- **An appendix with**
  - a link (e.g., to a dropbox folder) with the actual code and data. This code should reproduce the quantitative results in the PDF (e.g., the tables and figures).
  - the code as text.

Please keep your solution private (including the PDF and code) so that we can use the exercises in future years.

[1] See https://afajof.org/wp-content/uploads/files/afa_code_of_professional_con.pdf.

Exercise 1: Beta-Dollar Neutral Portfolio Construction

This exercise will help you think about factors, optimization, regressions, and projections. A factor can be made beta-neutral (i.e., no exposure to the overall market), dollar-neutral (no notional exposure, that is, an equal amount of money in, respectively, risky long positions and risky short positions), or both – denoted beta-dollar neutral. To explore beta-dollar neutrality further, suppose that the vector $y \in \mathbb{R}^N$ consists of long-short portfolio weights chosen with your favorite method (e.g., a decile sort), without regard to beta- or dollar-neutrality. Specifically, y is the portfolio at time t and it varies over time, but we just write y (rather than y_t) for simplicity, noting that the factor's excess return next time period is $y'r_{t+1}$. The vector of all stocks' current market betas is denoted β, and we use the notation $\vec{1} \in \mathbb{R}^N$ for a vector of ones.

1. Interpret the following problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^{N\times 1}} \ & (x-y)'(x-y) \\ \text{s.t. } & x'\vec{1} = 0 \\ & x'\beta = 0 \end{aligned}$$

   In other words, what is the interpretation of the solution, x, to this problem?

2. Rewrite and solve the problem using $B = (\vec{1}, \beta) \in \mathbb{R}^{N\times 2}$,

$$\begin{aligned} \min_{x\in\mathbb{R}^{N\times 1}} \ & (x-y)'(x-y) \\ \text{s.t. } & x'B = 0 \end{aligned}$$

3. Suppose that y is the strategy given by

$$y = a\vec{1} - \beta = B\begin{pmatrix} a \\ -1 \end{pmatrix}$$

   This strategy y is a form of betting-against-beta strategy – explain why and provide intuition for this strategy.

   When betas are one on average across stocks, $1 = \bar{\beta} := \frac{1}{N}\vec{1}'\beta$, interpret the meaning of $a = 1$ and $a = \beta'\beta/N$, respectively.

   For any choice of a, what is the corresponding beta-dollar-neutral version, x, of this trading strategy (i.e., the solution to part 2 in this case)?

4. Consider the cross-sectional regression of y on B:

$$y = B\theta + \varepsilon$$

   What is the OLS estimate of the regression coefficient, $\hat{\theta} \in \mathbb{R}^{2\times 1}$? What is the regression residual, $\hat{\varepsilon} = y - B\hat{\theta}$, and how does is it relate to the solution to 2?

Interpret this connection. Recall that $B\hat{\theta}$ is the projection of y on the span of B while $\hat{\varepsilon} = y - B\hat{\theta}$ is the projection on the orthogonal complement.

5. Reflect on the types of strategies that can, and cannot, be made beta-dollar neutral.

Exercise 2: Construct Value Factors

The goal of this exercise is to implement and analyze different value factors.

Value Factors

Value factors buy “cheap” stocks while shorting “expensive” ones. There are many different ways to implement a value factor depending on the definition of cheap/expensive (the “characteristic”), the investment universe, the rebalancing frequency, the portfolio weighting scheme, and so on. A standard value factor is that of Fama and French (1993), which you will implement in this exercise. The methodology has changed slightly since the original paper, and we will try to replicate the version available from Kenneth French’s data library. The construction of the underlying book-to-market characteristic is described here. For this exercise, you can rely on CRSP for constructing market equity, but you should supplement book equity data from Compustat, with the Moody’s data used in Davis et al. (2000). Finally, you will also work with the closely related value factors from Jensen et al. (2022).

WRDS Data

Empirical asset pricing researchers heavily use data from WRDS. The most common way to access data from WRDS is via the web interface. However, this way of accessing WRDS makes it difficult for other researchers to reproduce your dataset and only gives access to a subset of the data available on WRDS.

A more powerful approach is to work directly on the WRDS server via the WRDS Cloud. Currently, the cloud supports SAS, Python, R, and Stata. For the dataset used in Jensen et al. (2022), we relied on SAS and the corresponding IDE SAS Studio for data preparation. For this exercise, you may use whatever language you prefer, but you must download all WRDS data via the cloud and not via the web interface.

JKP Data

To solve this exercise, you can get inspiration from the source code used in Jensen et al. (2022), which you can find at `https://github.com/bkelly-lab/ReplicationCrisis`. However, the construction of the book-to-market equity characteristic in Jensen et al. (2022) differs from the Fama-French method, so we recommend that you write your data extracting code from scratch. For this exercise you also need to download the US book-to-market equity factors (equal-weighted, value-weighted, capped-value-weighted) from JKPfactors.com.

FF Data

Download the Fama-French value factor as a comparison (HML^{FF}). You can find all their data at `https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`, and you should use the monthly value factor that is part of the “Fama/French 3 Factors” in their “U.S. Research Returns Data.” The book equity used in Davis et al. (2000) is also available from this website.

Questions:

1. **Compute the characteristic.** Use US annual accounting data from Compustat and Moody's and monthly price data from CRSP to create the book-to-market ratio that underlies the Fama-French HML factor. Compute the book-to-market breakpoints (the 30 and 70 percentile) used for the HML factor. Make sure to compute the breakpoints on the longest samples possible, meaning that the first set of breakpoint should be from 1926. Describe your approach and plot how the breakpoints vary over time. *Optional: Add the Fama-French breakpoints for comparison.*

2. **Generate data set of stock returns.** Create a data set of monthly return data, describe your approach. Compute the standard deviation of returns each month, and plot the resulting series over time.

3. **Generate factor returns.** Combining the characteristic data and the return data, create a monthly return series of the HML factor in the US. The first return observation should be from July 1926. Describe your approach, plot the cumulative return of your factor alongside HML^{FF}, and report the Pearson correlation between the monthly returns of the two factors.

4. **Compare factor returns.** Compare and contrast the HML factor you have created with the US Book-to-Market equity factor from JKPfactors.com (both the equal-weighted, value-weighted, capped-value-weighted). Discuss the differences.

5. **Extra question, not required.** Do the same using another country, say Denmark. For this, you need to make assumptions on how to adapt the Fama-French definitions to international data.

Exercise 3: Factor Replication Analysis

The goal of this exercise is to analyze replicability of common factors from the academic literature. Replication rate means the fraction of factors where the coefficient of interest, average excess return or alpha, is positive and the two-sided p-value is statistically significantly positive at the 5% level. We require that the coefficient is positive because all factors from JKPfactors.com are signed such that the factor has a positive return according to the original study. As such, a factor is only replicated if it has a significantly positive return.

Questions:

The first three questions ask you to compute replication rates without multiple-testing adjustments.

1. **Excess returns.** Download all 153 US factors from the JKP data at a monthly frequency. Compute the average excess return of each factor and its OLS t-statistic. Compute the replication rate separately for equal-weighted (ew), value-weighted (vw), and capped-value-weighted (vw-cap) factors. I.e., compute three numbers.

2. **Originally significant factors.** Do the same as in question 1, but only for factors that were significant in the original study. For the list of these factors, download "Factor Details.xlsx" from `https://github.com/bkelly-lab/ReplicationCrisis/blob/master/GlobalFactors` and use the factors where the column "Significance" is 1. From here onwards, use only these factors in the computation of replication rates.

3. **Alphas.** Compute the CAPM alpha, its OLS t-statistic, and p-values. Compute the replication rate, again separately for ew, vw, and vw-cap. Download the most recent market returns from the link mentioned in the README from `https://github.com/bkelly-lab/ReplicationCrisis`.[2]

The next two questions ask you to use frequentist multiple-testing adjustments.

4. **Bonferroni.** Implement the Bonferroni adjustment to the alpha p-values and report the corresponding replication rates.

5. **Benjamini-Hochberg.** Implement the Benjamini-Hochberg adjustment to the alpha p-values and report the corresponding replication rates.

For the last two questions, use data from 1972 to 2021 to have a balanced panel. Next, write the return of factor i as

$$f_t^i = \alpha^i + \beta^i r_t^m + \epsilon_t^i,$$

and scale all market-adjusted returns, $f_t^i - \beta^i r_t^m = \hat{\alpha}^i + \hat{\epsilon}_t^i$, to have a monthly volatility of $0.1/\sqrt{12}$. Assume that the true alphas have the following hierarchical structure:

$$\alpha^i = \alpha^0 + c^j + w^i,$$

[2] Specifically, download the "market_returns.csv" file from this link: `https://www.dropbox.com/sh/zvrnbfg6u8ugo8o/AABugE3vXglTg-tr32Wb9-hHa?dl=0`.

where $\alpha^0 = 0$ is the component common to all factors, $c^j \sim N(0, \tau_c^2)$ is a cluster specific alpha, and $w^i \sim N(0, \tau_w^2)$ is a factor specific alpha. Assume that β^i is known and equal to its sample counterpart, that is, the OLS beta. Further, assume that the vector of idiosyncratic returns are distributed as $\epsilon_t \sim N(0, \Sigma)$. When estimating a large variance-covariance matrix such as Σ, we always need to ensure that it is non-singular to avoid dividing by zero when we compute Σ^{-1}. More broadly, we must ensure that the smallest eigenvalues are not too close to zero or too small relative to the largest eigenvalues (this ratio is called the matrix's condition number). To overcome this issue, we assume that the factor correlation matrix has a block structure. Specifically, we assume that the correlation between two factors, depend only on their cluster membership as defined in Jensen et al. (2022). The cluster memberships are available from `https://github.com/bkelly-lab/ReplicationCrisis/blob/master/GlobalFactors/Cluster%20Labels.csv`.[3]

6. **Block covariance matrix.** Separately for ew, vw, and vw-cap, estimate a 13 × 13 matrix where each entry is the average correlation between pairs of market-adjusted factor returns (i.e., $f_t^i - \beta^i r_t^m = \hat{\alpha}^i + \hat{\epsilon}_t^i$), where one factor is from the row cluster and the other is from the column cluster. For example, the intersection between the Value and Momentum cluster has the average correlation of pairs where one factor is from the Value cluster, and the other is from the Momentum cluster. Similarly, the diagonal for Value is the average correlation among factors belonging to the Value cluster. Show the cluster matrix for vw-cap factors.

   Build the corresponding 119 × 119 block correlation matrix, C^{block}, for all market-adjusted factor returns as follows. The diagonal consist of ones. Any off-diagonal element is the relevant entry in the cluster correlations that you just estimated.

   Finally, re-construct the covariance matrix as $\Sigma^{\text{block}} = \text{diag}(\sigma) C^{\text{block}} \text{diag}(\sigma)$, where $\text{diag}(\sigma)$ is a matrix with the sample volatilities (which should each be $0.1/\sqrt{12}$ if you scaled correctly) in the diagonal and zeros elsewhere.

7. **Empirical Bayes prior.** Separately for ew, vw, and vw-cap factors, use maximum likelihood to find the most likely value of τ_c and τ_w given the observed data and Σ^{block} from the previous question. Report your estimated values of τ_c and τ_w.

8. **Bayesian Posterior.** Based on Σ^{block}, τ_c and τ_w's from the previous questions, estimate the posterior distribution of alpha according to (24) and (25) in Jensen et al. (2022). Report the proportion of the factors where the posterior probability of a positive alpha is above 97.5% for each weighting scheme.

[3] For futher description of this approach, see appendix B in Jensen et al. (2022).

Exercise 4: High-Dimensional Return Prediction

The goal of this exercise is to implement three different models to predict stock returns in the UK, ranging from standard regression based predictions to machine learning.

Background

The equity factor literature is built around the concept of a linear low-dimensional return prediction framework. An equity factor is most often based on *one* characteristic, which corresponds to a univariate prediction problem. More broadly, the standard approach of predicting returns (or other variables of interest) using an OLS linear regression with a limited set of explanatory variables has the advantages of being simple to implement and providing easily interpretable regression coefficients based on the linear structure. The risk of overfitting is thus reduced by restricting ex ante the number of explanatory variables and by imposing a linear structure.

A recent trend in financial economics is to use machine learning methods to predict stock returns. Machine learning models are designed to handle many predictors and potentially allow non-linear dependence on these predictors while controlling the problem of overfitting. See Israel et al. (2020) for a discussion of machine learning in finance.

In this exercise, we will use the data from Jensen et al. (2022) (JKP) to predict international stock returns, similar to models analyzed for US stocks in Gu et al. (2020). Specifically, your task is to implement an OLS regression, a ridge regression, and a random forest to predict stock returns in the UK.

Data: Returns and Characteristics of Stocks in the UK

The return data from JKP is generated via the source code in `https://github.com/bkelly-lab/ReplicationCrisis/tree/master/GlobalFactors`, but you can simply download the data via this Dropbox link.

The input to your models should be the 21 characteristics called: `be_me`, `ret_12_1`, `market_equity`, `ret_1_0`, `rvol_252d`, `beta_252d`, `qmj_safety`, `rmax1_21d`, `chcsho_12m`, `ni_me`, `eq_dur`, `ret_60_12`, `ope_be`, `gp_at`, `ebit_sale`, `at_gr1`, `sale_gr1`, `at_be`, `cash_at`, `age`, `z_score`,
or the subset of the first 3 predictors (book-to-market, 12-month momentum, market equity). The models should be fitted based on in-sample return data from the beginning of the sample and until the end of 2011. The remaining return data (from the beginning of 2012 until the end of 2021) is the test set (or holdout data) used evaluate the performance of these models. If you need a validation set, use return data from the beginning of 2005 to the end of 2011. The objective is to predict excess returns, $r_{i,t}$, well out-of-sample. Note that it is perfectly acceptable to use publically available packages such as `ranger` in R or `scikit-learn` in Python.

Questions

1. **Basic data preparation.**[4] Take the following steps to prepare the data:

   (a) Exclude data before 1991-12-31.

   (b) Exclude observations with missing market equity in month t and missing return in month $t+1$.

   (c) Exclude observation with more than 5 out of the 21 characteristics missing.

   (d) Exclude nano caps (`size_grp='nano'`).

   (e) Choose some way to handle missing characteristics, e.g., by replacing them with the cross-sectional (i.e. within month) median.

   (f) *Optional:* Standardize the characteristics, e.g., by subtracting the mean and dividing by the standard deviation (but remember to be aware of look-ahead bias).

   Plot the number of available stocks over time.

2. **Fit each model using the training data.** Predict next month's excess return over the training period using only the 3 characteristics. Specifically, perform this analysis using the following three models:

   (a) **An OLS regression.**

   $$r_{i,t+1} = \beta' s_{i,t} + \varepsilon_{i,t+1} \tag{1}$$

   You can pool across stocks and time periods or, alternatively, use a Fama-MacBeth regression (a cross-sectional regression in each time period, and then take the time-series average of these).

   Report the estimated coefficients and a measure of uncertainty in these estimate (standard error or confidence interval).

   (b) **A ridge regression.**

   - For each ridge parameter λ, the ridge regression coefficient β_λ minimizes the objective:

   $$\min_{\beta} \sum_{i,t} (r_{i,t+1} - \beta' s_{i,t})^2 + \lambda \beta' \beta \tag{2}$$

   - To choose λ, split your training data into a training and validation subset,
   - Choose a range of lambda, suitable to your data, to search over.[5] For each λ, estimate the ridge regression over the training set, generating a regression coefficient $\hat{\beta}_\lambda$.

[4] You should not exclude more observations, but otherwise feel free to modify the data further, if think it will improve performance on the prediction task.

[5] Use the in-sample data to find an appropriate range for lambda.

- Use the estimated parameter to predict the return over the validation period, and pick the λ that has the lowest mean-squared error:

$$\min_{\lambda} \frac{1}{\#\text{observations}} \sum_{i,t \in \{\text{validation set}\}} (r_{i,t+1} - \hat{\beta}_{\lambda}' s_{i,t})^2 \tag{3}$$

  (A more sophisticated method is to use cross-validation, say 3-fold cross-validation. This means that you also use the first 1/3 data as validation sets, and the rest as training, and, likewise use the second 1/3 of the data as validation set. In this way, you have a non-cheating prediction for each return in the in-sample period, and you can compute the mean-squared error over the full in-sample period.)

- Based on the optimal ridge parameter, $\hat{\lambda}$, re-estimate the ridge regression coefficient using the entire in-sample data (training plus validation set).

Plot the mean squared error on the validation set, as a function of λ.

(c) **A random forests.**

- Fit 18 different random forest models to the training data that precedes the validation period. The hyper-parameters for the models are the 18 combinations of $P \in \{\frac{1}{3}, \frac{2}{3}, 1\}$, $D \in \{1, 2, 3\}$, and $M \in \{1, 10.000\}$, where P is the fraction of features to use for each tree,[6] D is the maximum tree depth, and M is the minimum number of samples in a leaf node.[7] Set the number of trees to 500 and the bootstrap sample size to 50% of the total observations available to the model. Make sure to specify a seed for the random number generator, such that your results can be reproduced.
- Use the estimated models to predict returns over the validation period, and pick the set of hyper-parameters with the lowest mean-squared error:

$$\min_{P,D,M} \frac{1}{\#\text{observations}} \sum_{i,t \in \{\text{validation set}\}} (r_{i,t+1} - \hat{f}^{RF}(s_{i,t}, P, D, M))^2 \tag{4}$$

- Based on the optimal hyper-parameters, re-estimate the random forest using the entire in-sample data (training plus validation set).

Plot the mean-squared error on the validation set, as a function of each of the three hyper-parameters i.e. show three different plots.

3. **In-sample differences.** For each of the three models, report the in-sample R^2_{is} computed as

$$R^2_{is} = 1 - \frac{\sum_{i,t \in \{\text{train}\}} (r_{i,t} - \hat{f}(s_{i,t-1}))^2}{\sum_{i,t \in \{\text{train}\}} (r_{i,t} - \bar{r}_{\text{train}})^2} \tag{5}$$

[6] That is, choose either 1, 2, or 3 and 7, 14, or 21 features when you use 3 or 21 features respectively.

[7] M is called `min.node.size` in `ranger` and `min_samples_leaf` in `scikit-learn`.

where $\bar{r}_{\text{train}} = \frac{1}{\#\text{observations}} \sum_{i,t \in \{\text{train}\}} r_{i,t}$ is the average return in the training set i.e. from the start of the sample to the end of 2011. Discuss the differences across models.

4. **Feature importance.** Show one measure of feature importance for each model, and discuss the differences.

5. **Out-of-sample performance.** Compute the performance of each model on the test set. Specifically, report the following:

   (a) The out-of-sample R^2_{oos} computed as

   $$R^2_{oos} = 1 - \frac{\sum_{i,t \in \{\text{test}\}} (r_{i,t} - f(s_{i,t-1}))^2}{\sum_{i,t \in \{\text{test}\}} (r_{i,t} - \bar{r}_{\text{train}})^2}$$

   where $\bar{r}_{\text{train}}$ is the average return over the training period as in (5). Note that Gu et al. (2020) replace $\bar{r}_{\text{train}}$ with $c = 0$.

   (b) Each month, sort stocks into five portfolios based on their predicted return. Compute the monthly value-weighted returns of these portfolios, as well as the 5-minus-1 long-short portfolio. Report the average excess returns, its t-statistic, the CAPM alpha, its t-statistic, the Sharpe ratio, and the information ratio (alpha devided by residual volatility) for all these portfolios.

   Discuss differences across models.

6. **Performance with more predictors.** Repeat the steps in question 2 to 5, but now using all 21 predictors.

Exercise 5: Research Proposal

The goal of this exercise is to get you started doing your own research. You should come up with an idea for a research paper and hand in two pages consisting of the following:

Page 1. The title of your paper, your name, and an abstract of at most 100 words. (It is a good exercise to write an abstract before you do all the hard work to see if there is possible scenario where something exciting comes out of the project, to get your ideas flowing, and to remind you that the project should end up as a paper with a main result. Plus, writing an abstract is simply a good way to train being very precise, clear, and concise.)

Page 2. A figure with a self-contained caption illustrating the main idea, plus a brief discussion and possibly some references. The figure can be made using real data, or, if this is not yet feasible, you can draw a figure by hand as you imagine it might look when you get the real data.

If you have an interesting figure illustrating a novel idea, an abstract, and a title, then you might be surprised how close you could be to having a real paper. To make this into a real paper, you need to check that the result is robust, consider the economic implications, and explain the result in writing. In the end, a paper is simply 6+ tables with a significant p-value that shows a novel result; or a theoretical result, possibly with an empirical test with fewer tables. It all comes down to a p-value and/or a q.e.d. That said, once you find a good idea and start making progress, work really hard to ensure that the execution is first rate. You may also find some inspiration on "How to Succeed in Academia or Have Fun Trying" here:

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3972340

References

Davis, J. L., E. F. Fama, and K. R. French (2000). Characteristics, covariances, and average returns: 1929 to 1997. *The Journal of Finance 55*(1), 389–406.

Fama, E. F. and K. R. French (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics 33*(1), 3–56.

Gu, S., B. Kelly, and D. Xiu (2020). Empirical Asset Pricing via Machine Learning. *Review of Financial Studies 33*(5), 2223–2273.

Israel, R., B. T. Kelly, and T. Moskowitz (2020). Can Machines Learn Finance? *Journal of Investment Management 18*(2), 23–36.

Jensen, T. I., B. T. Kelly, and L. H. Pedersen (2022). Is There A Replication Crisis In Finance? *Journal of Finance, Forthcoming*.