The zip file associated to this document contains all Python codes as well as the data necessary to reproduce all figures and tables in the main text of the paper "Lassance, N., Vanderveken, R. & Vrins, F. Optimal Portfolio Size under Parameter Uncertianty. *Journal of Financial and Quantitative Analysis*, forthcoming.". It is composed of two sections: Replication Inputs and codes.

1. Replication inputs

This section presents the primary datasets used in the paper's empirical analysis, along with additional computed quantities referenced throughout. These include outputs from simulation exercises and various pre-computed elements, such as degrees of freedom or random permutations, used to accelerate the empirical computations.

1.1. Datasets

The paper considers seven different datasets in the empirical analysis, which are listed in Table 1 of the main text. We provide in the zip file seven .csv files of monthly asset returns of size $T \times N$ where T is the number of observations, and N the number of assets.

1. *100SBTM_m_08_2023.csv* contains the dataset of 100 portfolios formed on size and book-to-market from July 1926 through August 2023. We only use 96 portfolios because four portfolios have missing data. We refer to this dataset as 96S-BM in the paper.

2. *108CHA.txt* contains the dataset of 108 characteristic portfolios built on long and short legs of 54 characteristics in Lassance and Martín-Utrera (2023). It spans from September 1966 to December 2020, and we refer to this dataset as 108CHA in the paper.

3. *100SOP_m_08_2023.csv* contains the dataset of 100 portfolios sorted on size and operating profitability from July 1963 through August 2023. We refer to this dataset as 100S-OP in the paper.

4. *48IND_46ANOM_m_12_2013.csv* contains the dataset of 48 industry portfolios and 46 characteristic portfolios built on long and short legs of 23 characteristics in Novy-Marx and Velikov (2015). It spans from July 1973 to December 2013, and we refer to this dataset as 94IN-NV in the paper.

5. *25sbtm_ind48_opinv25_mom10_m_12_2022.csv* contains the dataset of 47 industry portfolios, 25 portfolios sorted on size, and book-to-market, 25 portfolios sorted on operating profitability and investment and 10 portfolios sorted on momentum. It spans from July 1963 to December 2022, and we refer to this dataset as 107IN-CHA in the paper.

6. *48IND_16ANOM_OPINV25_10MOM_m_12_2013* contains the dataset of 47 industry portfolios, 25 portfolios sorted on operating profitability and investment, 10 portfolios sorted on momentum, and 16 characteristic portfolios built on long and short legs of

eight low-turnover characteristics in Novy-Marx and Velikov (2015). It spans from July 1963 to December 2013, and we refer to this dataset as 98IN-CHA-NV in the paper.

7. *25sbtm_ind48_opinv25_mom10_m_12_2022.csv* contains the dataset of 47 industry portfolios, 25 portfolios sorted on size, and book-to-market, 25 portfolios sorted on operating profitability and investment and 10 portfolios sorted on momentum. It spans from July 1973 to December 2022, and we refer to this dataset as 107IN-CHA in the paper.

8. *RET.csv* contains the dataset of 235 US stocks which we use across 100 datasets of 100 stocks, randomly drawn from the total pool of 235 stocks. It spans from January 1998 to March 2022, and we refer to this dataset as 100STO in the paper.

The industry portfolios, the portfolios sorted on size and book-to-market, the portfolios sorted on operating profitability and investment, and the portfolios sorted on momentum are obtained from Kenneth French's website (`https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html`). The Fama-French three factors can also be obtained from the same website, and are provided in order to compute excess asset returns, in the file *FF_3factors_m_08_2023.csv*. The seventh dataset is collected from the Center for Research in Security Prices (CRSP).

1.2. Additional inputs

The remaining inputs can be split into two main categories: simulation data and inputs for the empirical analysis.

1.2.1. Simulation data

The files in this section store the results of the various simulation exercises implemented throughout Section 4 of the main text. Three main files are in the zip file:

1. *Nstar_[strat]_ellasy_6.csv* files. These .csv files store the estimated portfolio sizes used in Figure 4. These quantities are computed in the Python code *rep_fig3.py*. *[strat]* is a placeholder that changes according to the considered portfolio strategy: *kz* for the 2F strategy of Kan and Zhou (2007), *smv* for the SMV portfolio, *smvgmv* for the 3FGMV rule, *smvew* for the 3FEW rule, *smvl1* for the norm-constrained SMV-ST rule, *smv_aetw* for the SMV-HT rule, and *wba* for the F+A rule.

2. *oosu_eachsim_empcorr_6_[strat]_nstar_ellasy.csv* files. These .csv files store the out-of-sample utility (oosu) delivered by each strategy (2F, SMV, 3FGMV, 3FEW) for each portfolio size N and for each simulation of the simulations exercise detailed in Section 4.2 of the main text. *empcorr_6* indicates that these are the results for $\nu = 6$ and $\mathbf{\Sigma}_M = \hat{\mathbf{\Sigma}}_{96}$.

3. *mean_oosu_sim_empcorr_6_[strat].csv* files. These .csv files aggregate the results in the *eachsim* files by averaging over all simulations.

1.2.2. Inputs for the empirical analysis

Because the empirical analysis is both large-scale and computationally intensive, several quantities are pre-computed and provided as inputs to the main code. This substantially speeds up the generation of results. These pre-computed objects are stored in the following files:

1. Estimates of $\tilde{\kappa}^s_{N,1}$ and $\tilde{\kappa}^s_{N,2}$ for each time step and each dataset. The estimation procedure is described in the Appendix. The estimates of $\tilde{\kappa}^s_{N,1}$ are stored in *etastock_[dataset]_[permutation].csv*, and the estimates of $\tilde{\kappa}^s_{N,2}$ in *upsilonstock_[dataset]_[permutation].csv*. The placeholder *[permutation]* identifies which of the random draws of 100 portfolios is used for the seventh dataset.

2. Estimates of the optimal portfolio sizes $\hat{N}^\star$ for the main rules considered in the empirical analysis, again for each time step and each dataset. These are stored in the *nstar_portfolios_[dataset]_sample_[permutation].csv* files.

In addition, due to the structure of the empirical analysis, numerous random permutations of assets must be generated. These permutations arise naturally for the 100STO dataset, as well as in the implementation of the Rand and Best θ^2_N selection rules. To ensure reproducibility, all asset permutations are saved in separate .csv files:

1. *selmat_[dataset]_[permutation].csv*: For each dataset and time step, these files contain 1000 random selections of assets. They serve two purposes: (i) implementation of the Best θ^2_N selection rule, and (ii) implementation of the Rand selection rule. Files for the STO100 dataset are not included due to their size but can be provided upon request.

2. *idxpermutmax_95Q_[dataset]_selection.csv*: For each dataset and time step, these files store the index (within the corresponding *selmat* file) of the asset selection achieving the 95% quantile of the estimated θ^2_N values. This is used for the Best θ^2_N selection rule.

3. *selection_mat_STO235_100.csv*: These files contain the random selections of assets used specifically in the empirical analysis of the 100STO dataset.

2. Codes

Several Python codes are provided in this package. For each code, the user must change the *your_path* string variable to indicate the path of the replication inputs provided with this replication package.

1. *rep_figs_1_2_4.py*. This code replicates Figures 1, and 2.

2. *rep_fig_3.py*. This code replicates Figure 3.

3. *rep_fig_4.py*. This code replicates Figure 4.

4. *rep_fig_5.py*. This code replicates Figure 5.

5. *rep_fig_6.py*. This code replicates Figure 6.

6. *rep_fig_7.py*. This code replicates Figure 7.

7. *rep_fig_8.py*. This code replicates Figure 8.

8. *rep_package_empirical.py*. This code replicates the empirical analysis implemented in Section 5 of the main text. Running it as such yields the results for the 2F strategy for the 100S-BM dataset, for all selection rules. The results are generated similarly for the other portfolio strategies and datasets, and can be generated by changing the strategies and/or datasets in the code.

References

Kan, R. and Zhou, G. (2007), "Optimal portfolio choice with parameter uncertainty." *Journal of Financial and Quantitative Analysis*, 42(3):621–656.

Lassance, N. and Martín-Utrera, A. (2023), "Sentiment-based portfolios." *SSRN working paper*, (2023).

Novy-Marx, R. and Velikov, M. (2015), "A taxonomy of anomalies and their trading costs." *Review of Financial Studies*, 29(1):104–147.