Students: Andrea Contenta
Professor: Guillaume Monarcha

Assignment

Master 203
Financial Econometrics 2

1. Through principal component analysis of the European equity returns, extract the K main latent factors (principal components) F_k of the European equity market, i.e. the core factors that drive the returns of European equities.

We apply Principal Component Analysis (ie PCA) to standardised stock returns to get i principal components (ie PCs). Standardising returns is a necessary step to avoid scale effect when aggregating various returns.

We can analyse our first results by looking at Figure 1 and Figure 2. We correctly see that most of the variance is captured from few components (ie, more than 40% with just a component), whereas the marginal increment is smaller and smaller if we move rightwards. We also (coherently) see that the R^2 increases as we add up PCs. Though, we know that R^2 is not a good measure of goodness-of-fit if we are comparing models with a different number of explanatory variables (as we are doing here), as its value increases mechanically. Instead, we could use some information criteria that balance the trade-off between adding new variables (ie, minimising bias) and loosing degrees of freedom (ie, increasing variance).

Some authors have pointed out the possibility of using R^2 as a decision criterion for PCA. Darolles and Mero[1] provide a method to estimate a target R^2 (and a noise-to-signal ratio) that will be used as a threshold, through a Monte Carlo simulation. This methodology avoids one of the main drawbacks of IC: they do not have a global minimum, so we will never know if we are choosing the "best" model.

2. For each stock i: identify the k_i^* significant core factors and estimate their associated exposures $\hat{b}_{k,i}$ from the following linear model: $r_{i,t} = \alpha_i + \sum_{k=1}^{k^*} b_{k,i} F_{k,i} + \varepsilon_t$

Various methods have been proposed in the literature to select the most significant explanatory variables. There are two main approaches:

- focusing on an econometric and machine learning method that does the selection itself (ie, Least Absolute Shrinkage and Selection Operator - LASSO);
- focusing on an information criterion that aims at comparing various models.

When applying PCA, the second method looks preferable. This is because model selection methods are particularly designed to cope with **compensation effects** that arise when explanatory variables are significantly correlated (ie, they show multicollinearity). Instead, PCs' correlation matrix is by construction set to (approximately) zero. So, there is not any compensation effect to fix when using PCs as explanatory variables (ie, as factors as we are doing here).

[1] S. Darolles, G. Mero, *Hedge Fund Returns and Factor Models: A Cross-Sectional Approach*, Bankers, Markets Investors, n$^{\text{o}}$ 112, May-June 2011, 34-53

We will focus on information criteria (ie, IC) set by Bai-Ng[2]. Python allows us to compute three different version of these IC, so that we may not end up with a clear decision path. As a general rule, we aim at minimising the their value, as they are a direct function of the cross-sectional average residual variance. When different IC would select different models, we should rank the three IC with regard to their **penalisation factor** - which is what makes they differ. Through application of Bai-Ng IC, we can affirm that the optimal model is the one with four PCs.

```
The number of optimal PCs that minimize Bai-Ng information criteria is 4
```

One of the main drawbacks of this approach is that we have set a number of criteria by looking at the whole cross-section returns. This implies that the first four components should have a good explanatory power if regressed against each return. In reality, we see that the i R^2 took from the regression of each return against the fours PCs are dramatically different. This has a significant impact both for in-sample fit and for out-of-sample predictions. To cope with this problem, we could set a target R^2 and selecting the PCs that help us reach that level (but again, R^2 is not good when comparing models with a different number of explanatory variables, so we should use an adjusted version). By doing so, each stock should be explained by a different subset of PCs. Despite this drawback, in the end, we opt for Bai-Ng criteria for the sake of simplicity.

3. Compute the weights of the K equity portfolios designed to replicate the K core equity factors.

Factor mimicking portfolios (ie, FMPs) aim at building portfolios of (investable) assets that replicate the dynamic of un-investable assets. Here, we will try to build a portfolio of stocks that replicate their first four principal components. This will allow investors have an exposure to an "invisible" asset, with the same logic developed by Jurczenko-Teiletche[3]. We use the function `port_minvol_ro`[4] to solve the minimisation problem while setting the desired return to a certain level (ie, the PC). In fact, using a generic `port_minvol` function would be wrong in this context, since the optimized weights would be totally unrelated to the PCs that we are trying to replicate.

Figures 5 to 8 show the four FMPs against the respective mimicking factor. We see that the model can undoubtedly optimised to better fit the factor. The main drawback of the current analysis is that we opted for a dynamic OLS regression. When this is used to estimate a dynamic multi-factor model, the results are biased for various reasons (ie, estimation error,

[2] J. Bai, S. Ng, *Determining the number of factors in approximate factor models*, Econometrica, Vol. 70, No. 1 (January, 2002), 191–221

[3] E. Jurczenko, J. Teiletche, *Macro Factor Mimicking Portfolios*, November 2019

[4] developed during the course

length of time interval, smoothing the coefficients). To solve this problem, we could opt for a Kalman filter. This would allow us to set a specific dynamic of factor exposures and to solve the bias embedded in dynamic OLS estimation. Also, Kalman filter would allow more reliable forecasts than OLS, as it is currently used in the asset management industry for **nowcasting**.

4. Estimate the alpha of these K portfolios against the market benchmark.

Estimating the alpha of a portfolio against only one index can be seen as practical application of the CAPM model. In this model, portfolio returns can be seen as a linear function of the market benchmark (minus the risk-free rate, ie equity risk premium) plus a constant. Here, we do not use any risk free rate since this can be arbitrary and time variant, but we manage to estimate consistently alphas and betas of FMPs. We just apply a linear regression model and then extrapolate the constant value.

The results are showed in figure 9: we coherently obtain some alphas around the value of zero, but this is not enough to affirm that there is a significant alpha. To do so, other methods have been explored in the literature. Kosowski *et al.*[5] propose a bootstrapped estimation that will result in a distribution of alphas rather than on an exact value. This is based on a two-step regression:

- the first one aims at estimating the factor exposures, at computing the residuals and get their empirical distribution;
- the second one aims at regressing the bootstrapped series of returns on the same factors, but setting alpha value to zero

Finding a significant constant value in the second regression would demonstrate the existence of the alpha.

5. Assess the impact of errors in the estimation of the covariance matrix $\hat{\Omega}$, on the estimated alphas of the K replicating portfolios. Compute their confidence intervals at the 95% level.

We have seen in the previous question that alphas are subject to estimation errors. This is mainly due to **normality assumption**: when this is violated, the results of a portfolio optimization problem dramatically change. This has a practical impact in the field of asset pricing since this is used to asses portfolio managers' skills.

[5] R. Kosowski, A. Timmermann, R. Wermers, and H. White, *Can Mutual Fund Stars Really Pick Stocks? New Evidence from a Bootstrap Analysis*, Sept 2005

To practically show the impact, we will adopt a bootstrap procedure. This is structured as follows:

- take a random sample of the initial sets of returns;
- compute resampled mean vector and covariance matrix;
- get the optimized FMPs weights using `port_minvol_ro` and setting the principal components as targeted return;
- compute returns through the optimized portfolio weights, and sum them to obtain bootstrapped portfolio returns;
- apply CAPM to the bootstrapped portfolio returns against the market benchmark to get bootstrapped alpha

We then plot the result in various ways to see the results. Most importantly, we obtain a distribution of bootstrapped alphas, and their relative boxplot in figure 11 and 12. They show that alphas do not behave as a gaussian random variable in reality. We see that there is a negative skew as the mean in concentrated at the right of zero. The boxplot also suggest that, even if the median value is almost the same over the four portfolios, the values can span greatly and there are also some outliers.

One of the main drawback of this analysis is the computational power. We simulated 50 steps to grasp the essential concepts, but more steps are needed to dive into the analysis.

We also report for the sake of curiosity other graphs that have been obtained over the course of the project. We see that the distribution of betas is drastically different: each of the four FMPs express a different mean, and a higher kurtosis. Figure 15 simulated the distribution of the estimation error on the first four components of an asset. Figure 10 proposes the same result but through a scatterplot.

To obtain confidence intervals of alpha at 95%, we consider the top and bottom 2.5% quantiles, and we print a statement that reports the values:

```
The 95% confidence interval of global bootstrapped alphas is [-0.0024,0.0170].
The 95% confidence interval of the first bootstrapped alphas is [-0.0005,0.0210].
The 95% confidence interval of the second bootstrapped alphas is [-0.0003,0.0159].
The 95% confidence interval of the third bootstrapped alphas is [-0.0031,0.0168].
The 95% confidence interval of the fourth bootstrapped alphas is [0.0001,0.0165].
```

In fact, we can assess the distribution of the whole alphas, or each of the four FMPs' alphas distributions.

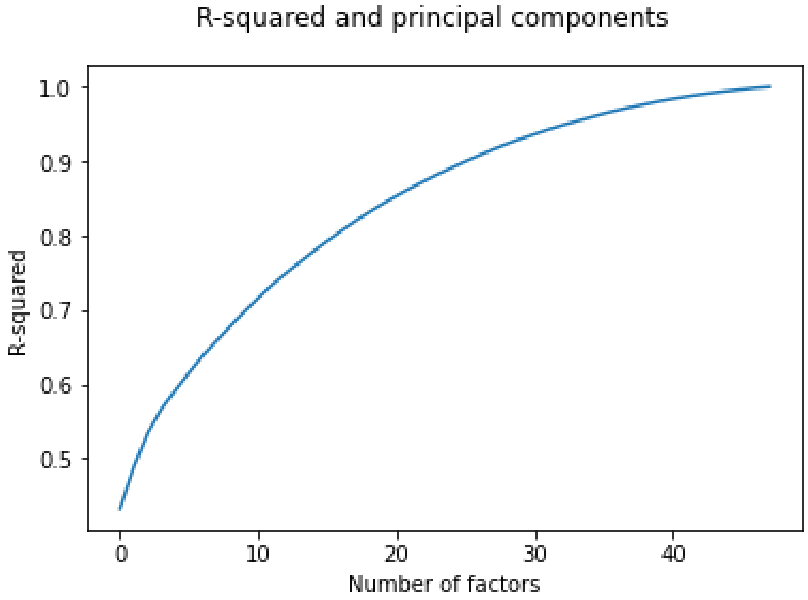

Figure 1: R^2 and principal components

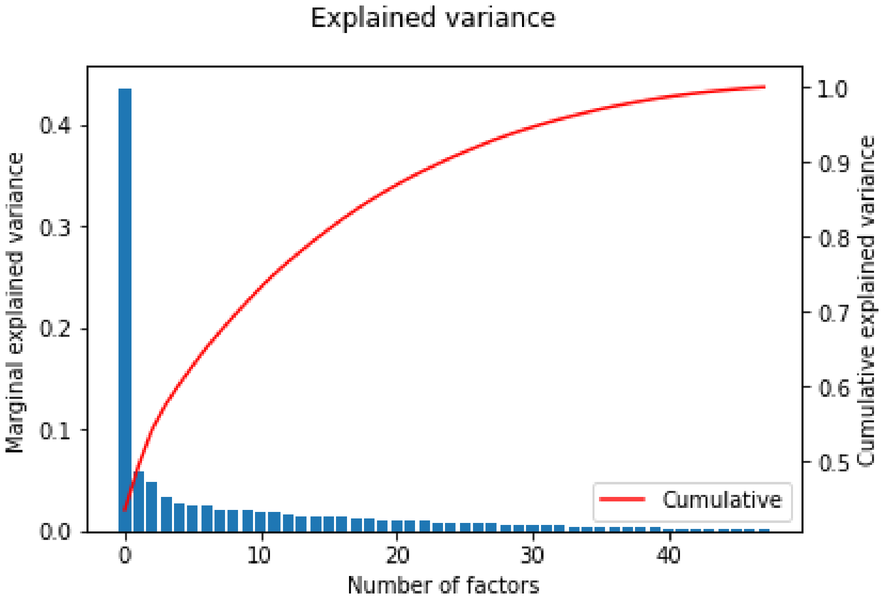

Figure 2: Explained (cumulative) variance of PCA models

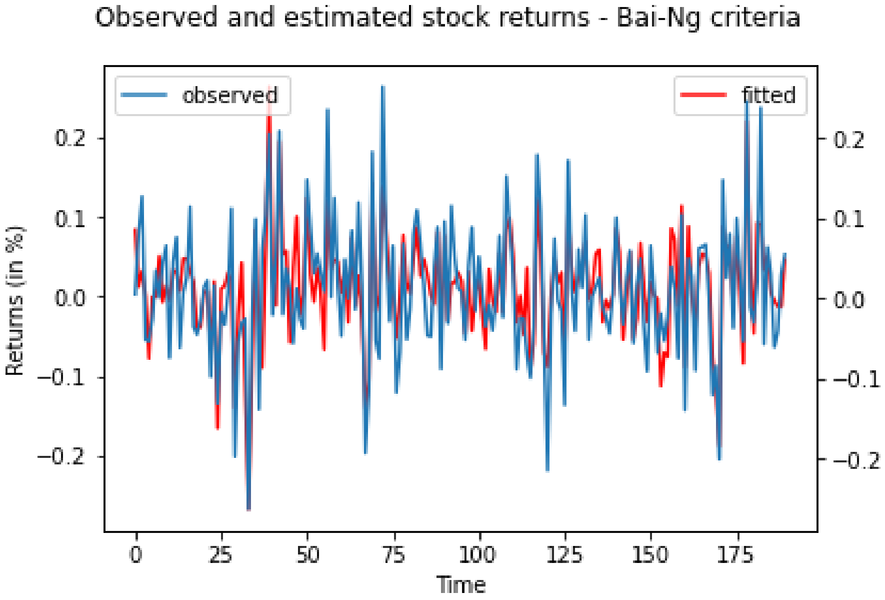

Figure 3: Example of a linear factor model

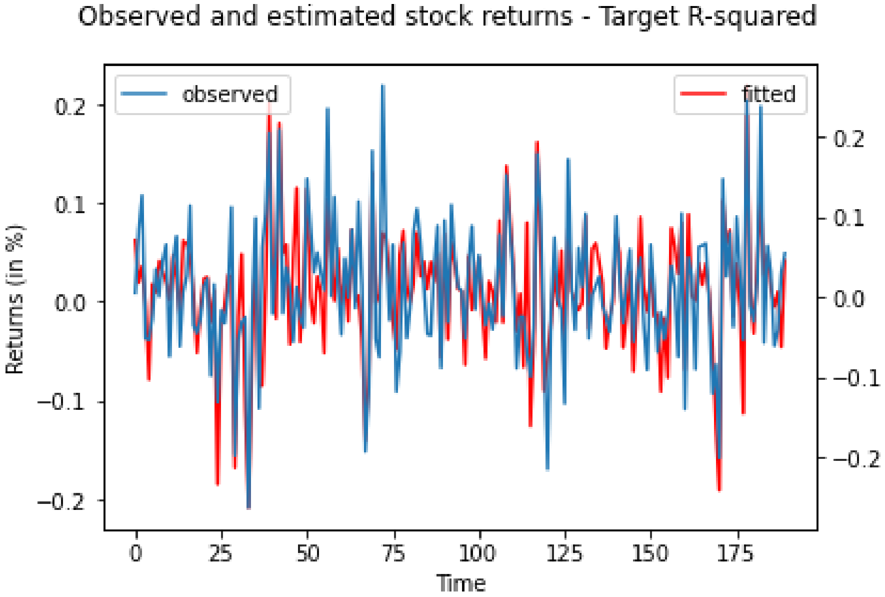

Figure 4: Example of a linear factor model

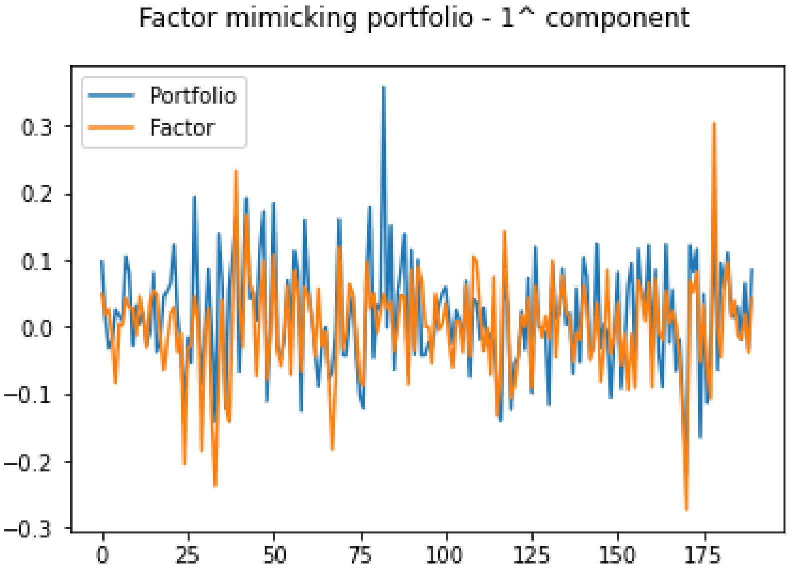

Figure 5: First factor mimicking portfolio

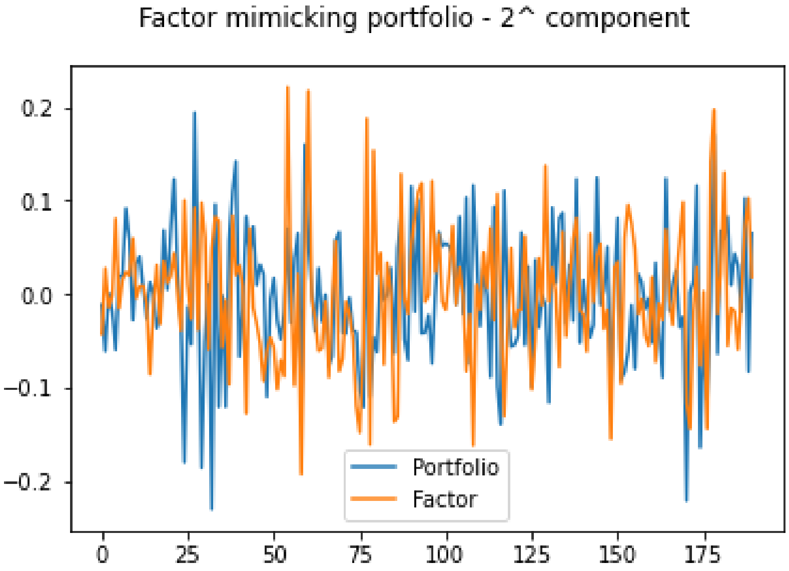

Figure 6: Second factor mimicking portfolio

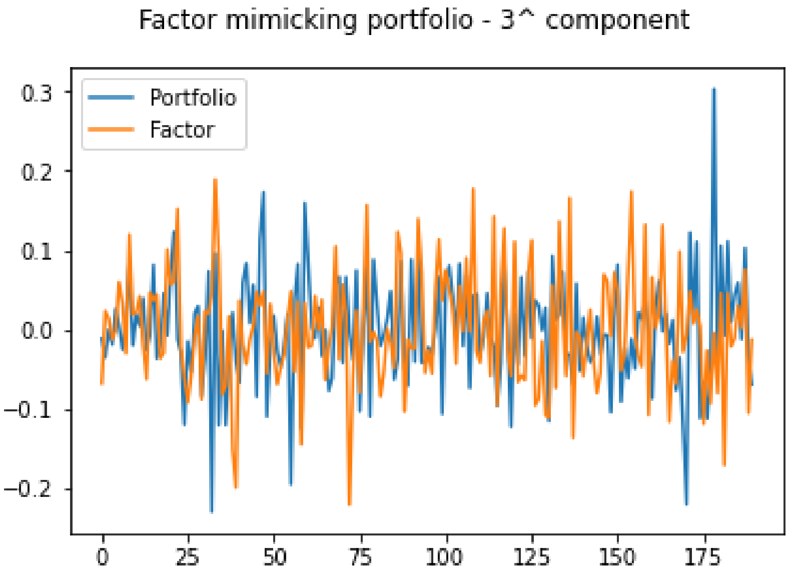

Figure 7: Third factor mimicking portfolio

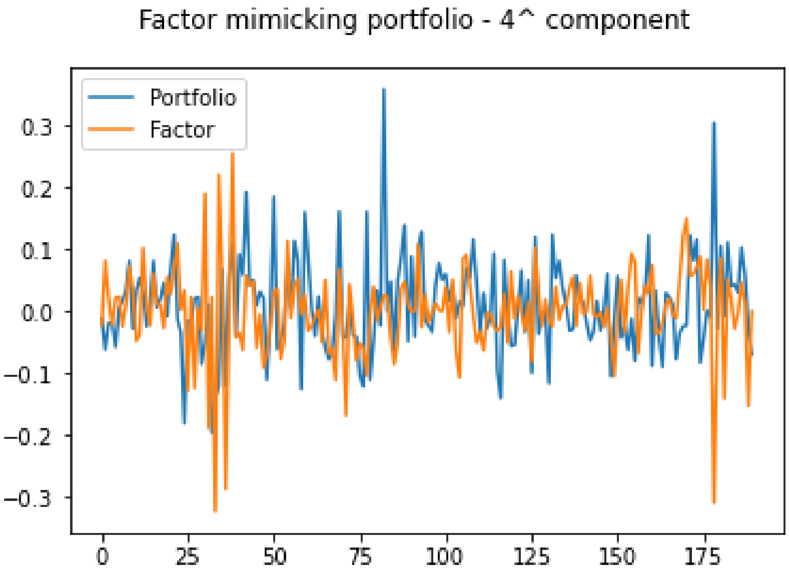

Figure 8: Fourth factor mimicking portfolio

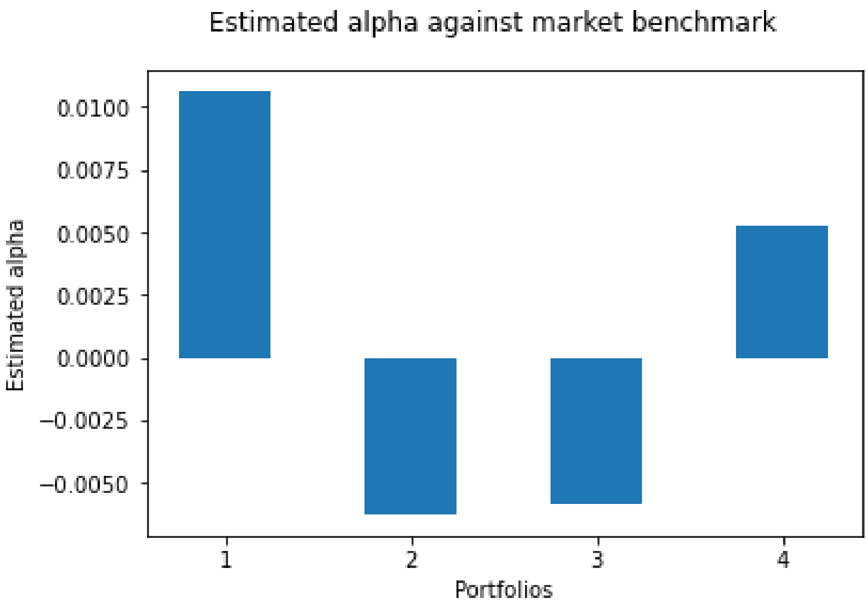

Figure 9: Estimated alphas of the four FMPs

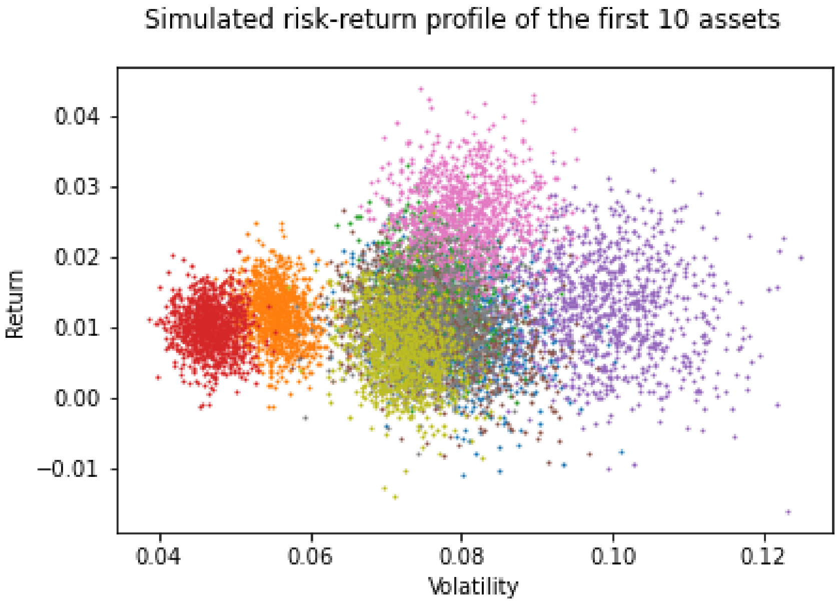

Figure 10: Simulated impact of estimation error on risk-return profile

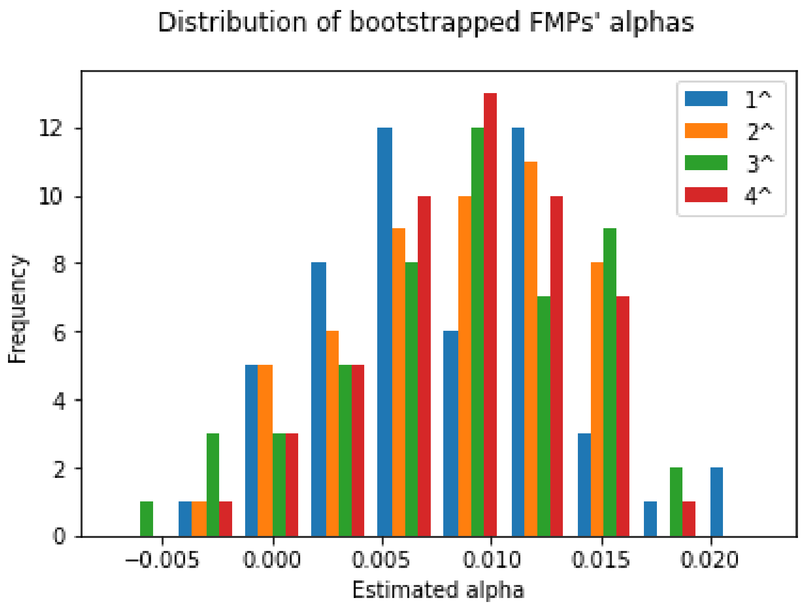

Figure 11: Distribution of bootstrapped FMPs' alphas

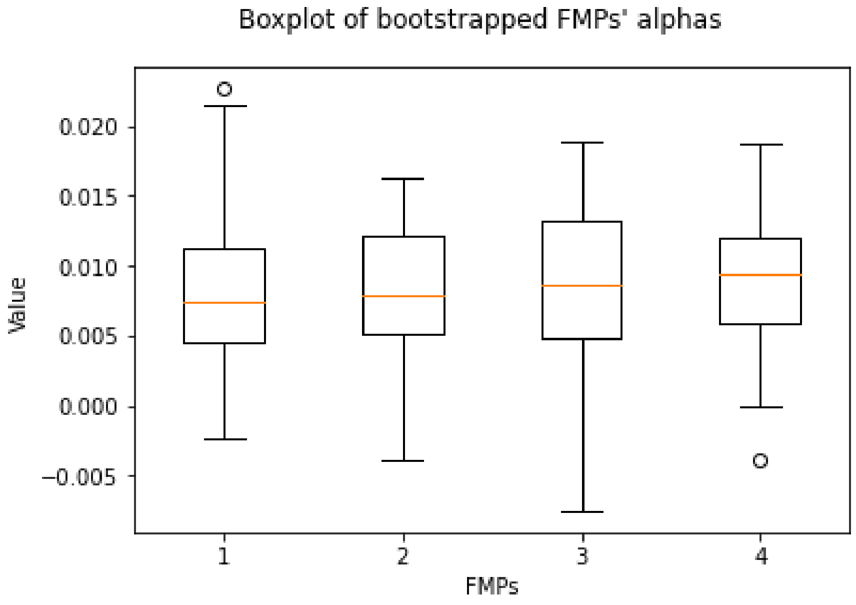

Figure 12: Boxplot of bootstrapped FMPs' alphas

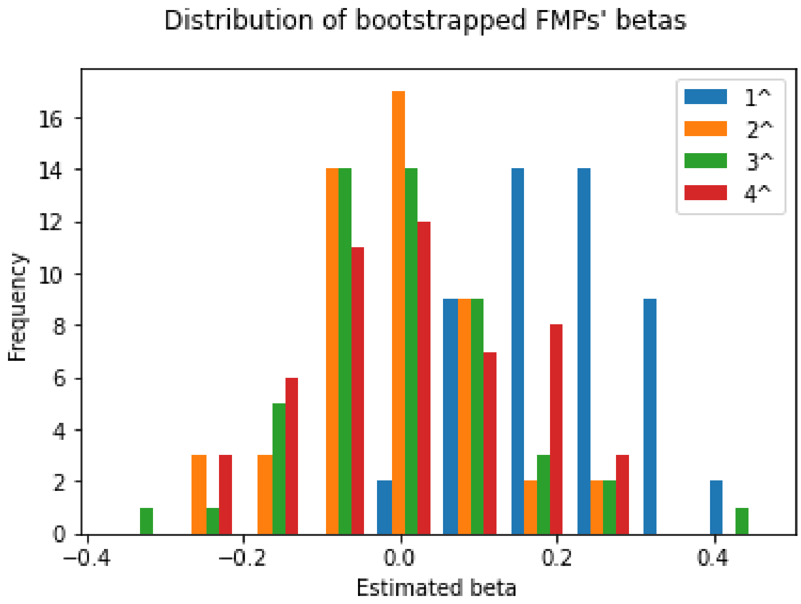

Figure 13: Distribution of bootstrapped FMPs' betas

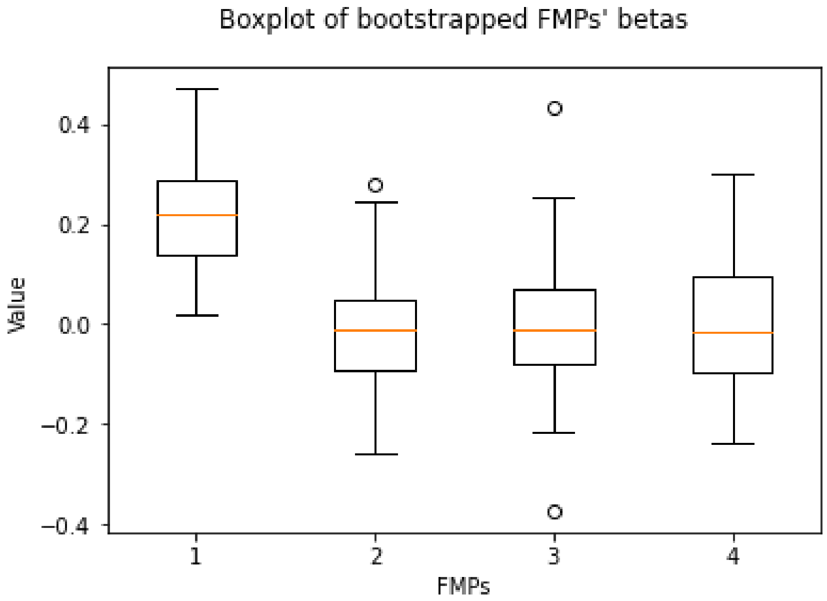

Figure 14: Boxplot of bootstrapped FMPs' betas

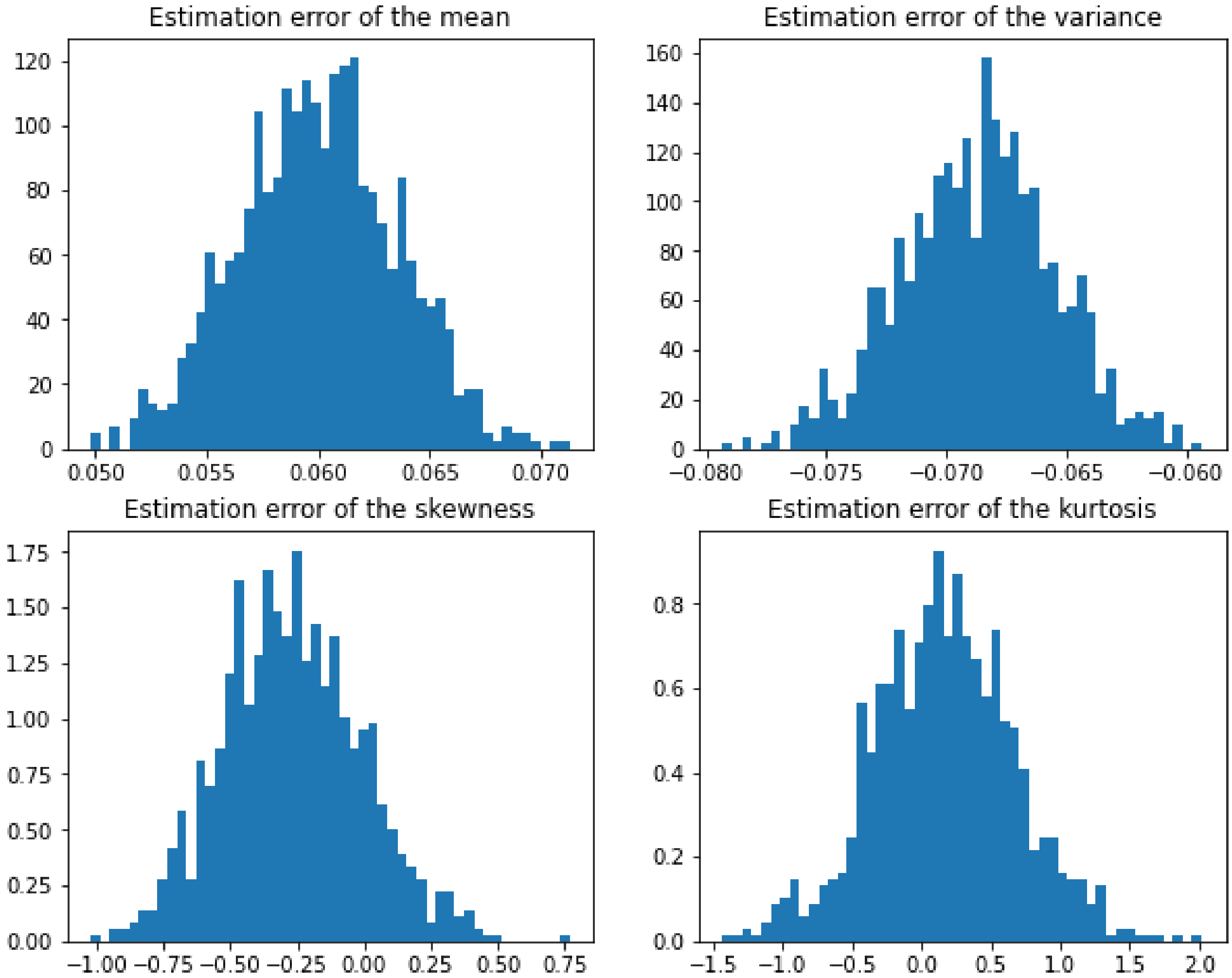

Figure 15: Simulated distribution of estimation error on a certain stock