Knockoffs for Trading US Equities

BSc Project Presentation

Zayn Shuman

Supervised by: Dr Arman Hassanniakalager

Overview of Presentation

1. Introduction
2. False Discovery Rate
3. Linear Modelling
4. Construct knockoffs
5. Asset Selection
6. Portfolio Construction
7. Q&A

Overview of Presentation

1. Introduction
2. False Discovery Rate
3. Linear Modelling
4. The Knockoff Filter
5. Asset Selection
6. Portfolio Construction
7. Q&A

Introduction

Question: How can we use statistical methods to select US equities capable of tracking the performance of a chosen Index?

Chosen indices: S&P 500, Russell 1000 and DJIA

US Equities: ~2000 US Equities traded on NYSE, AMEX, and Nasdaq exchanges

Overview of Presentation

1. Introduction
2. False Discovery Rate
3. Linear Modelling
4. The Knockoff Filter
5. Asset Selection
6. Portfolio Construction
7. Q&A

False Discovery Rate

Benjamini and Hochberg [1] introduced the False Discovery Rate (FDR) in 1995.

	Declared non-significant	**Declared significant**	**Total**
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
	$m - R$	R	m

Table 1: Testing m null hypotheses

Define: $Q = V/(V + S)$

V is the number of erroneously rejected hypotheses: $H_0\ is\ True, H_0\ rejected$
S is the number of correctly rejected hypotheses: $H_0\ is\ non - true, H_0\ rejected$

[1] Yoav Benjamini and Yosef Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: Journal of the Royal Statistical Society. Series B (Methodological) 57.1 (1995), pp. 289–300. issn: 00359246. url: `http://www.jstor.org/stable/2346101`

False Discovery Rate

The False Discovery Rate is the expected proportion of incorrectly rejected null hypotheses

$$FDR = \mathbb{E}(Q) = \mathbb{E}(V/\underbrace{(V + S)}) = \mathbb{E}(V/R)$$

$$\mathrm{V} + \mathrm{S} = 0 \;\Rightarrow Q = 0$$

False Discovery Rate

The False Discovery Rate is the expected proportion of incorrectly rejected null hypotheses

$$FDR = \mathbb{E}(Q) = \mathbb{E}(V/(V + S)) = \mathbb{E}(V/R)$$

Consider testing $H_1, H_2, H_3, \dots, H_m$

FDR Controlling Procedure

1. Each $H_1, H_2, H_3, \dots, H_m$ has a corresponding p-value $P_1, P_2, P_3, \dots, P_m$
2. Order the p-values $P_{(1)} \leq P_{(2)} \leq P_{(3)} \leq \cdots \leq P_{(m)}$ such that $P_{(j)}$ corresponds to null $H_{(j)}$
3. Let k be the largest i for which $P_{(i)} \leq \frac{i}{m} q^*$
4. Reject all $H_{(i)}: i = 1,2,3, \dots, k$

False Discovery Rate

Theorem 1

For independent test statistics and for any configuration of false null hypotheses, the FDR controlling procedure controls the FDR at q^*.

Overview of Presentation

1. Introduction
2. False Discovery Rate
3. Linear Modelling
4. The Knockoff Filter
5. Asset Selection
6. Portfolio Construction
7. Q&A

Linear Modelling

Barber and Candès [2] introduced the knockoff filter for statistical linear models.

Discover which features are truly associated with the response in a linear model

Guaranteed control of the FDR

Computation is cheap and construction does not require any new data

Method can work with a broad class of test statistics

[2] Rina Foygel Barber and Emmanuel J. Candès. “Controlling the false discovery rate via knockoffs”. In: The Annals of Statistics 43.5 (Oct. 2015). issn: 0090-5364. doi: 10.1214/15-aos1337. url: http://dx.doi.org/10.1214/15-AOS1337.

Linear Modelling

- Measuring returns from Friday to Friday, the first Friday of 2010 was on 2010-01-08 and the last Friday of 2019 was on 2019-12-27.
- Include only the largest 100 companies by market capitalisation

$$\boldsymbol{n} = 520$$
$$\boldsymbol{p} = 100$$

$$\overset{\boldsymbol{y}}{\begin{pmatrix} \textbf{Index} \\ \text{W1 Return} \\ \text{W2 Return} \\ \text{W3 Return} \\ \dots \\ \text{W}n\text{ Return} \end{pmatrix}} = \overset{\boldsymbol{X}}{\begin{pmatrix} \textbf{Equity } \boldsymbol{j} \\ \text{W1 Return} \\ \text{W2 Return} \\ \text{W3 Return} \\ \dots \\ \text{W}n\text{ Return} \end{pmatrix}} \overset{\boldsymbol{\beta}}{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \dots \\ \beta_p \end{pmatrix}} + \overset{\boldsymbol{z}}{\begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ \dots \\ z_n \end{pmatrix}}$$

Linear Modelling

Data Pre-processing

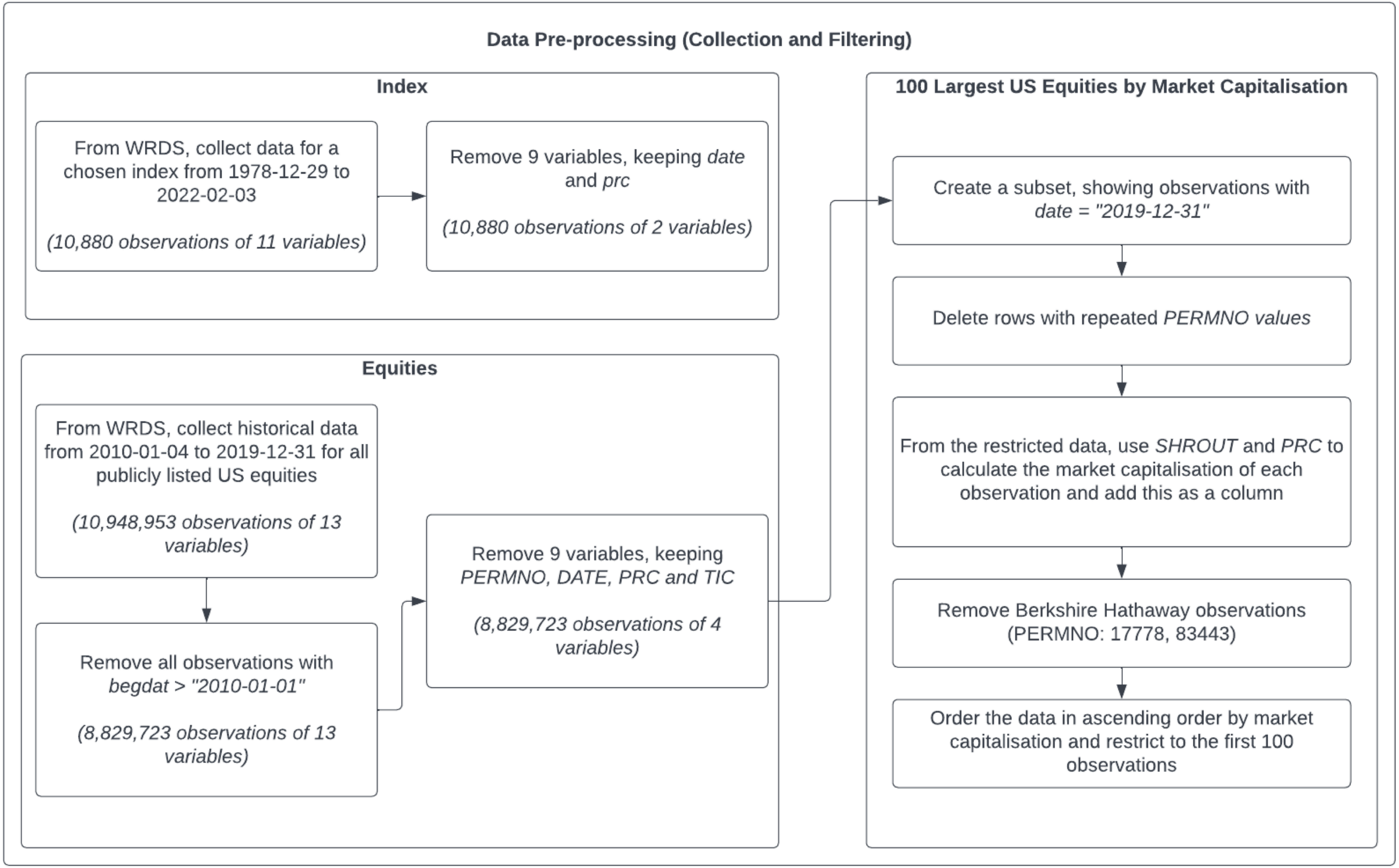
1. **PERMNO**
2. **BEGDAT**
3. **SHROUT**

Index dataframe

119,680 to 21,760 elements

Equities dataframe

142,336,389 to 35,318,892 elements


Data Pre-processing (Collection and Filtering)
Index
From WRDS, collect data for a chosen index from 1978-12-29 to 2022-02-03
(10,880 observations of 11 variables)
Remove 9 variables, keeping date and prc
(10,880 observations of 2 variables)
Equities
From WRDS, collect historical data from 2010-01-04 to 2019-12-31 for all publicly listed US equities
(10,948,953 observations of 13 variables)
Remove all observations with begdat > "2010-01-01"
(8,829,723 observations of 13 variables)
Remove 9 variables, keeping PERMNO, DATE, PRC and TIC
(8,829,723 observations of 4 variables)
100 Largest US Equities by Market Capitalisation
Create a subset, showing observations with date = "2019-12-31"
Delete rows with repeated PERMNO values
From the restricted data, use SHROUT and PRC to calculate the market capitalisation of each observation and add this as a column
Remove Berkshire Hathaway observations (PERMNO: 17778, 83443)
Order the data in ascending order by market capitalisation and restrict to the first 100 observations

Overview of Presentation

1. Introduction
2. False Discovery Rate
3. Linear Modelling
4. The Knockoff Filter
5. Asset Selection
6. Portfolio Construction
7. Q&A

The Knockoff Filter

Construct knockoffs

For each $\boldsymbol{X_j}$ in our design matrix, we construct a knockoff feature $\widetilde{\boldsymbol{X_j}}$

Calculate $\boldsymbol{\Sigma} = \boldsymbol{X}^T\boldsymbol{X}$, after normalising we want:

$$\widetilde{\boldsymbol{X}}^{\boldsymbol{T}}\widetilde{\boldsymbol{X}} = \boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{\Sigma}$$

$$\boldsymbol{X}^{\boldsymbol{T}}\widetilde{\boldsymbol{X}} = \boldsymbol{\Sigma} \; - \mathrm{diag}(\boldsymbol{s})$$

Comparing a feature to its knockoff:

$$\boldsymbol{X}_{\boldsymbol{j}}^{\boldsymbol{T}}\widetilde{\boldsymbol{X}_{\boldsymbol{j}}} = \Sigma_{\mathrm{jj}} \; - \mathrm{s}_{\mathrm{j}} = 1 \; - \mathrm{s}_{\mathrm{j}}$$

The Knockoff Filter

Construct knockoffs

Construction Strategy

$$\tilde{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{I} \ - \boldsymbol{\Sigma}^{-1}\text{diag}(\mathbf{s})) + \tilde{\boldsymbol{U}}\boldsymbol{C}$$

$\tilde{\boldsymbol{U}}$ is an $n \times p$ orthonormal matrix that is orthogonal to the span of the features $\boldsymbol{X}$

$$\boldsymbol{C}^{\mathrm{T}}\mathbf{C} = \mathbf{2}\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\ \boldsymbol{\Sigma}^{-1}\text{diag}(\mathbf{s})$$

SDP knockoffs

Minimise: $\Sigma(1 - s_j)$

Subject to: $0 \leq s_j \leq 1$

$\text{diag}(\mathbf{s}) \preccurlyeq 2\boldsymbol{\Sigma}$

$\Leftrightarrow$

$\boldsymbol{E}_{\boldsymbol{i}_{(\boldsymbol{r},\boldsymbol{c})}} = -1, 0 \text{ otherwise}$

Maximise: $tr\big(\boldsymbol{I}\ \text{diag}(\mathbf{s})\big)$

Subject to: $\Sigma s_i E_i \succcurlyeq -2\Sigma$

$tr\big(-E_i\text{diag}(\mathbf{s})\big) \leq 1$

$tr\big(E_i\ \text{diag}(\mathbf{s})\big) \leq 1$

The Knockoff Filter

Construct knockoffs

Construction Strategy

$$\tilde{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{\Sigma}^{-1}\text{diag}(\mathbf{s})) + \tilde{\boldsymbol{U}}\boldsymbol{C}$$

$\tilde{\boldsymbol{U}}$ is an $n \times p$ orthonormal matrix that is orthogonal to the span of the features $\boldsymbol{X}$

$$\boldsymbol{C}^{\mathrm{T}}\mathbf{C} = \mathbf{2}\text{diag}(\mathbf{s}) - \text{diag}(\mathbf{s})\,\boldsymbol{\Sigma}^{-1}\text{diag}(\mathbf{s})$$

Equi-correlated knockoffs

Maximise: $\boldsymbol{s}$

Subject to: $2\Sigma \succcurlyeq \text{diag}(\mathbf{s})$

$\boldsymbol{s} \geq \mathbf{0}$

Consider the constraint

$$2\Sigma \succcurlyeq \text{diag}(\mathbf{s}) \Leftrightarrow \lambda_{min}\big(2\Sigma - \text{diag}(\mathbf{s})\big) \geq \mathbf{0}$$
$$\Leftrightarrow 2\lambda_{min}(\Sigma) \geq s_j$$

Obvious solution: $s_j = 2\lambda_{min}(\Sigma)$

The Knockoff Filter

Calculate statistics for each pair of original and knockoff variables

We now wish to introduce the statistics W_j for each β_j

- Large positive values are evidence against the null hypothesis $H_0\colon\ \beta_j = 0$

Consider Tibshirani's Lasso model [3] as a method for constructing coefficient estimates

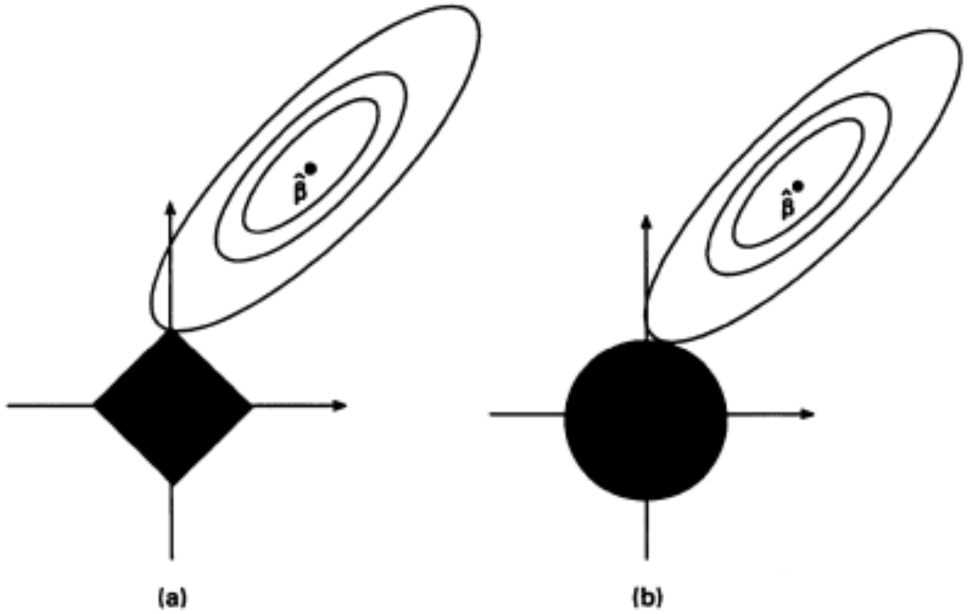
$$\widehat{\boldsymbol{\beta}}(\lambda) = argmin_{\boldsymbol{b}}\left\{\frac{1}{2}\underbrace{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}||_2^{\boldsymbol{2}}}_{\text{OLS Estimate}} + \boldsymbol{\lambda}\overbrace{||\boldsymbol{b}||_1}^{\text{Constraint: } \Sigma_j \beta_j \leq t}\right\}$$

[3] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: Journal of the Royal Statistical Society. Series B (Methodological) 58.1 (1996), pp. 267–288. issn: 00359246. url: `http://www.jstor.org/stable/2346178` (visited on 04/17/2022)

The Knockoff Filter

Calculate statistics for each pair of original and knockoff variables

$$\hat{\boldsymbol{\beta}}(\lambda) = argmin_{\boldsymbol{b}}\left\{\frac{1}{2}\underbrace{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}||_2^{\mathbf{2}}}_{\text{OLS Estimate}} + \overbrace{\boldsymbol{\lambda}||\boldsymbol{b}||_1}^{\text{Constraint: } \Sigma_j \beta_j \leq t}\right\}$$



Comparison between Lasso (a) and ridge (b) regression

[3] Robert Tibshirani. “Regression Shrinkage and Selection via the Lasso”. In: Journal of the Royal Statistical Society. Series B (Methodological) 58.1 (1996), pp. 267–288. issn: 00359246.
url: `http://www.jstor.org/stable/2346178` (visited on 04/17/2022)

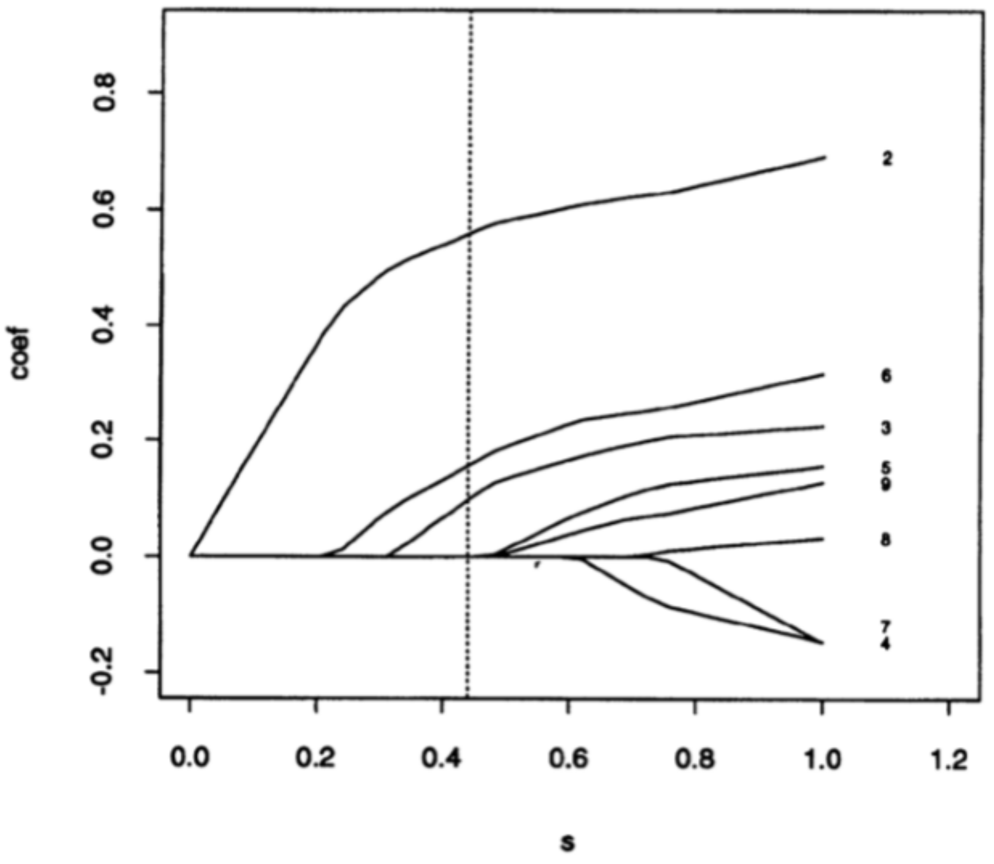
The Knockoff Filter

Calculate statistics for each pair of original and knockoff variables

$$Test\ statistic\ for\ feature\ j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$$

1. Apply the Lasso model to the augmented matrix $[\boldsymbol{X}, \widetilde{\boldsymbol{X}}]$
2. Construct a vector of test statistics $(Z_1, Z_2, \dots, Z_p, \widetilde{Z}_1, \widetilde{Z}_2, \dots, \widetilde{Z}_p)$
3. For each j, define W_j

$$W_j = Z_j \vee \widetilde{Z}_j \cdot \begin{cases} +1: Z_j > \widetilde{Z}_j \\ -1: Z_j < \widetilde{Z}_j \\ 0: Z_j = \widetilde{Z}_j \end{cases}$$



Lasso shrinkage of coefficients, $s = 1/\lambda$ *and* $coef = |\beta_j|$

The Knockoff Filter

Calculating a Threshold for the Statistics

We wish to select large positive W_j such that $W_j \geq t$ for some $t > 0$

Let $W = \{|W_j| : j = 1, \ldots, p\} / \{0\}$

For some target FDR q, define the following data-dependent threshold

$$T = \min\left\{t \in W : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\}$$

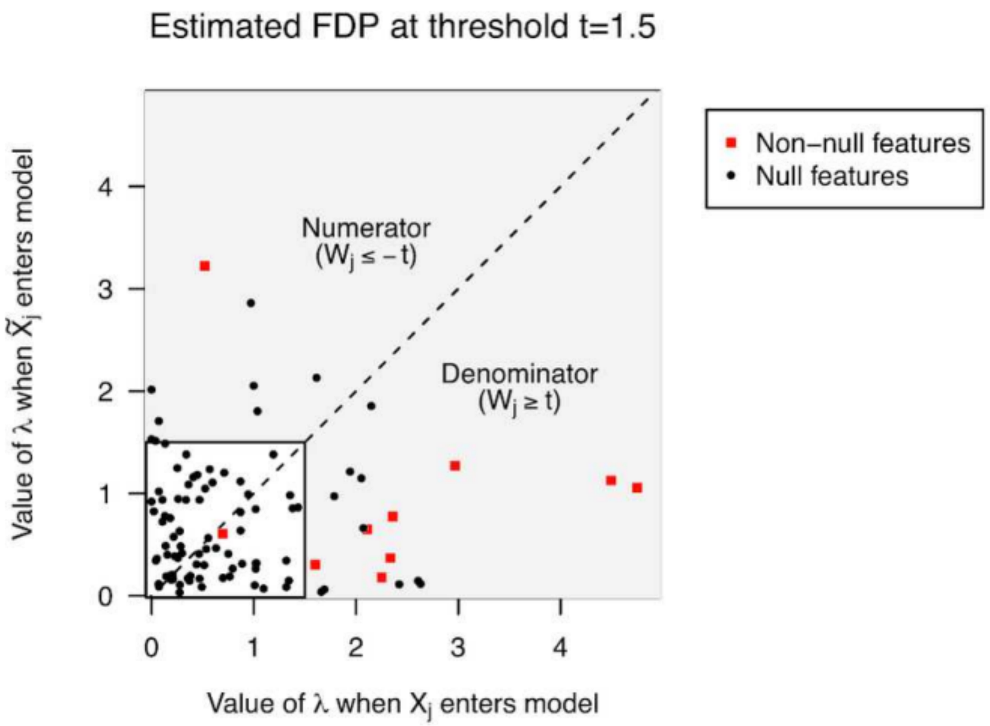
The Knockoff Filter

Calculating a Threshold for the Statistics

Data Dependent Threshold:

$$T = \min\left\{t \in W: \frac{\#\{j: W_j \leq -t\}}{\#\{j: W_j \geq t\} \vee 1} \leq q\right\}$$

- 9 features above the diagonal and 18 below: 9/18 estimates the FDR
- 8 true discoveries out of 18 selected features: 8/18 is the true FDR



Visualisation of the knockoff procedure, black points correspond to $\beta_j = 0$ and red points correspond to $\beta_j \neq 0$.

The Knockoff Filter

Theorems from Knockoffs

Define: $\hat{S} = \{j : W_j \geq T\}$

Theorem 2

The knockoff procedure controls a quantity nearly equal to the FDR in feature selection.

More Specifically:

$$\mathbb{E}\left[\# \frac{\{j\colon \beta_j = 0 \; and \; j \in \hat{S}\}}{\#\{j\colon j \in \hat{S}\} + q^{-1}}\right] \leq q$$

The Knockoff Filter

Theorems from Knockoffs

The knockoff+ procedure:

$$T' = \min\left\{t \in \mathcal{W}: \frac{1 + \#\{j: W_j \le -t\}}{\#\{j: W_j \ge t\} \vee 1} \le q\right\}$$

Define: $\hat{S} = \{j : W_j \ge T'\}$

Theorem 3

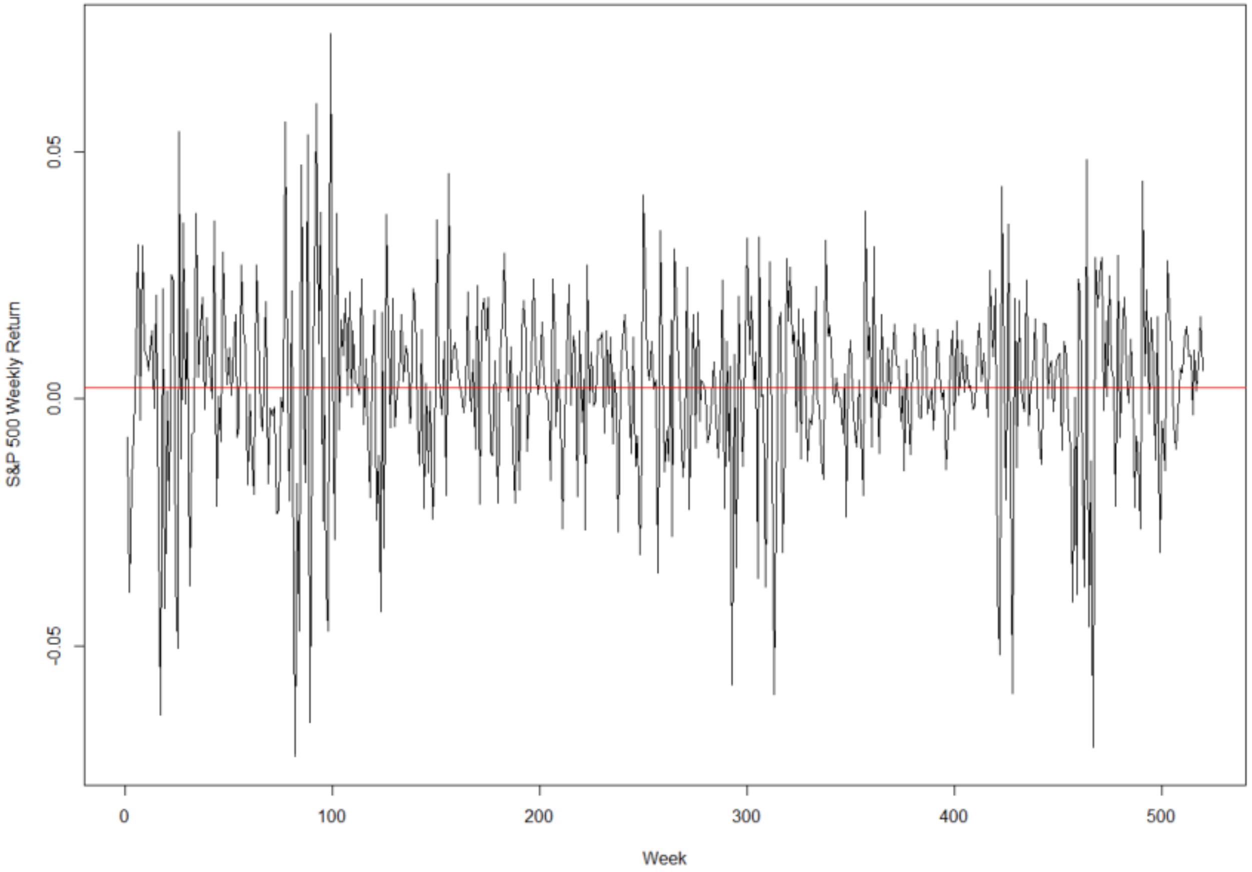
The knockoff+ procedure controls the FDR exactly in feature selection.

More Specifically:

$$\mathbb{E}\left[\# \frac{\{j: \beta_j = 0 \; and \; j \in \hat{S}\}}{\#\{j: j \in \hat{S}\} \vee 1}\right] \le q$$

Overview of Presentation

Asset Selection



Return series of the S&P 500 Index from 2010-01-15 to 2019-12-27

Mean weekly returns

0.002190

Observed variance

0.0003734

Asset Selection

- Target FDR of 0.05
- ‘knockoff’ package used in R
- 42 selected US equities

AAPL	MSFT	AMZN	JPM	MA	XOM	UNH
DIS	PFE	CMCSA	CSCO	PEP	C	ORCL
ADBE	NVDA	TMO	RTC	HON	TXN	DHR
SBUX	CVS	MO	USB	LOW	BKNG	MS
CAT	GS	MDLZ	FISV	ANTM	TFC	PROV
ISRG	PCS	SPGI	BSX	SCHQ	ITW	ECL

S&P 500 Knockoff Portfolio constituents

Overview of Presentation

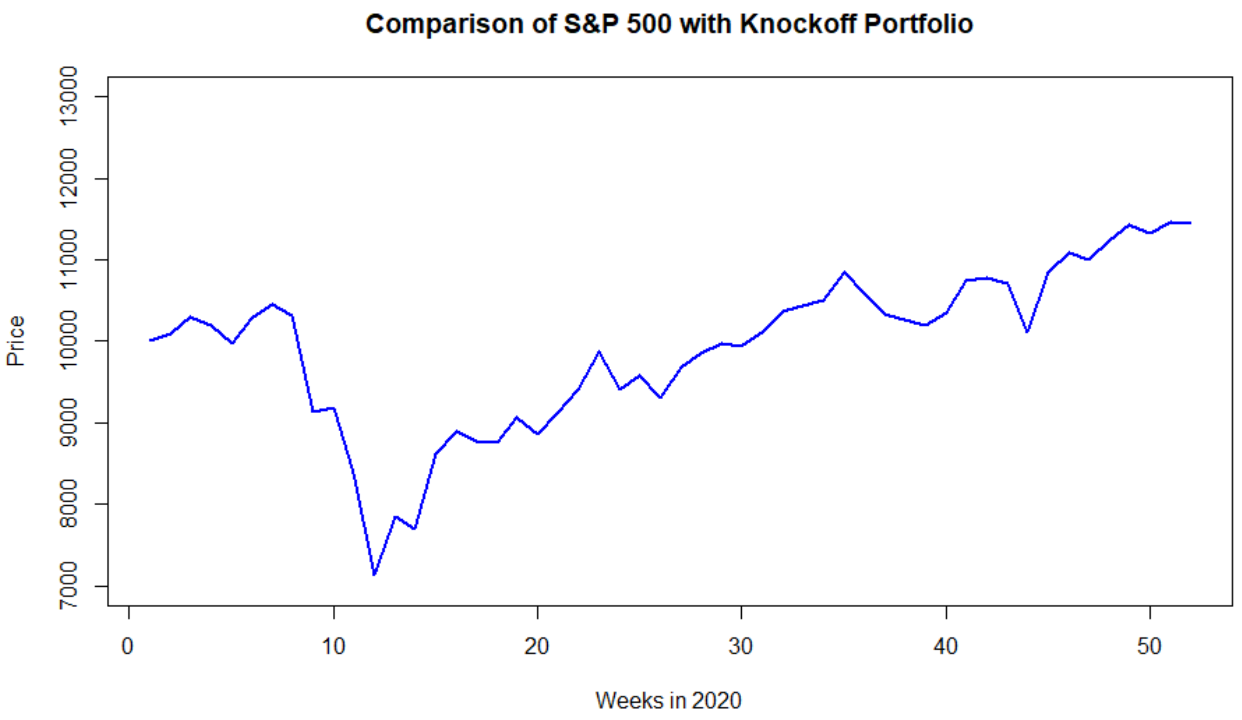
1. Introduction
2. False Discovery Rate
3. Linear Modelling
4. The Knockoff Filter
5. Asset Selection
6. Portfolio Construction
7. Q&A

Portfolio Construction

Weighting by Market Capitalisation

Replicate Index weightings

What would happen if I invested $10,000 on the first Friday of January 2020?

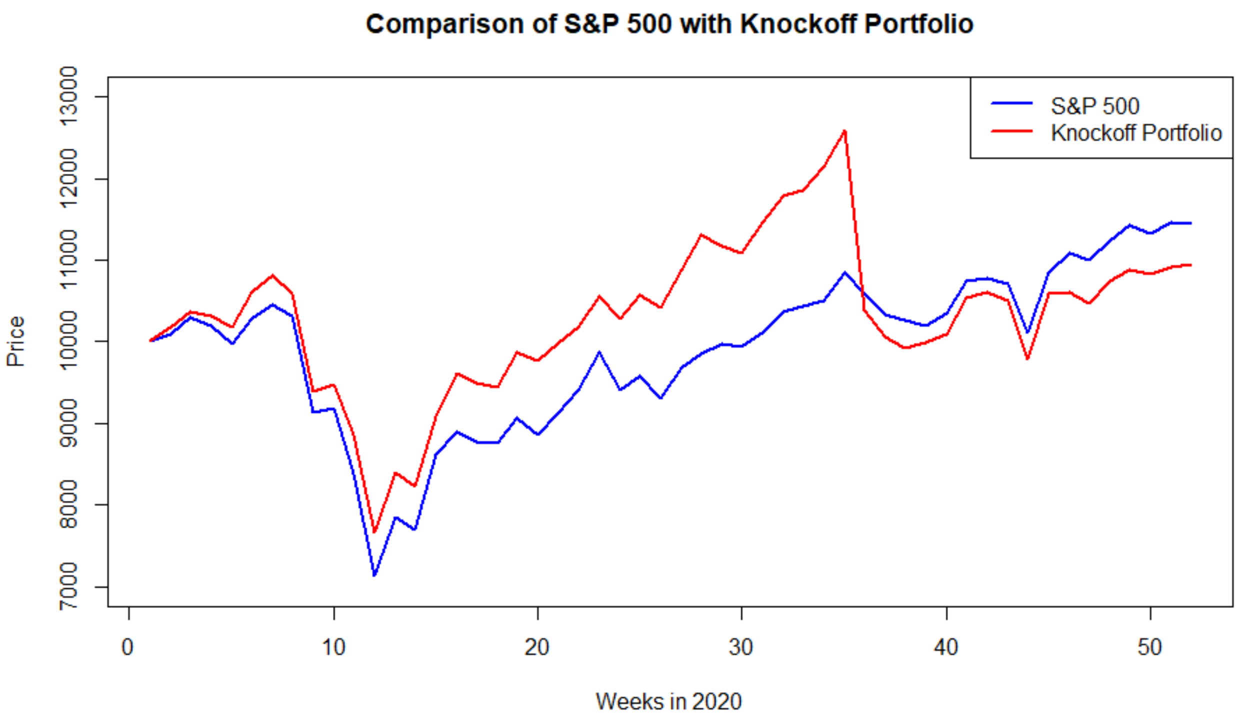


Portfolio Construction

Weighting by Market Capitalisation

Replicate Index weightings

What would happen if I invested $10,000 on the first Friday of January 2020?

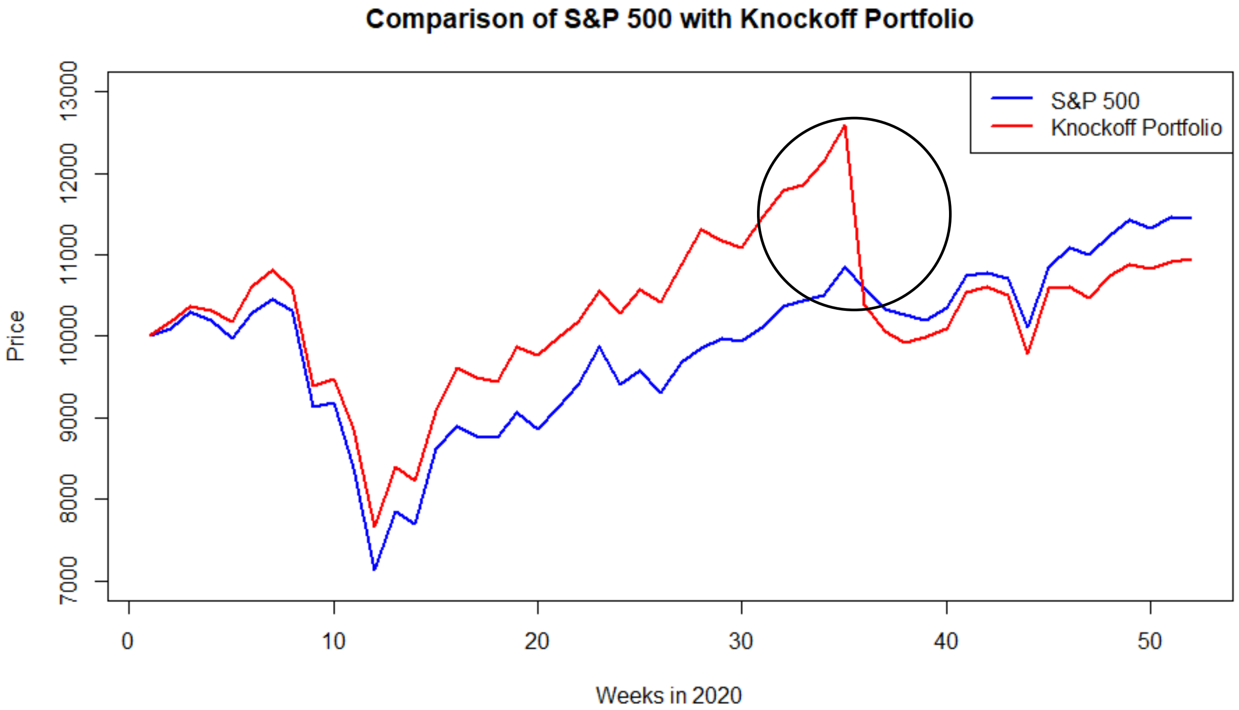


Portfolio Construction

Weighting by Market Capitalisation

Replicate Index weightings

What would happen if I invested $10,000 on the first Friday of January 2020?



Portfolio Construction

Modern Portfolio Theory

MPT assumes that investors are risk averse

Portfolio return variance: $\boldsymbol{w}^T \boldsymbol{\Sigma} \mathbf{w}$

Expected portfolio returns: $\boldsymbol{R}^T \boldsymbol{w}$

Covariance matrix of returns

Minimise: $\boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}$

Subject to: $\boldsymbol{R}^T \boldsymbol{w} = \mu^*$ → Target return

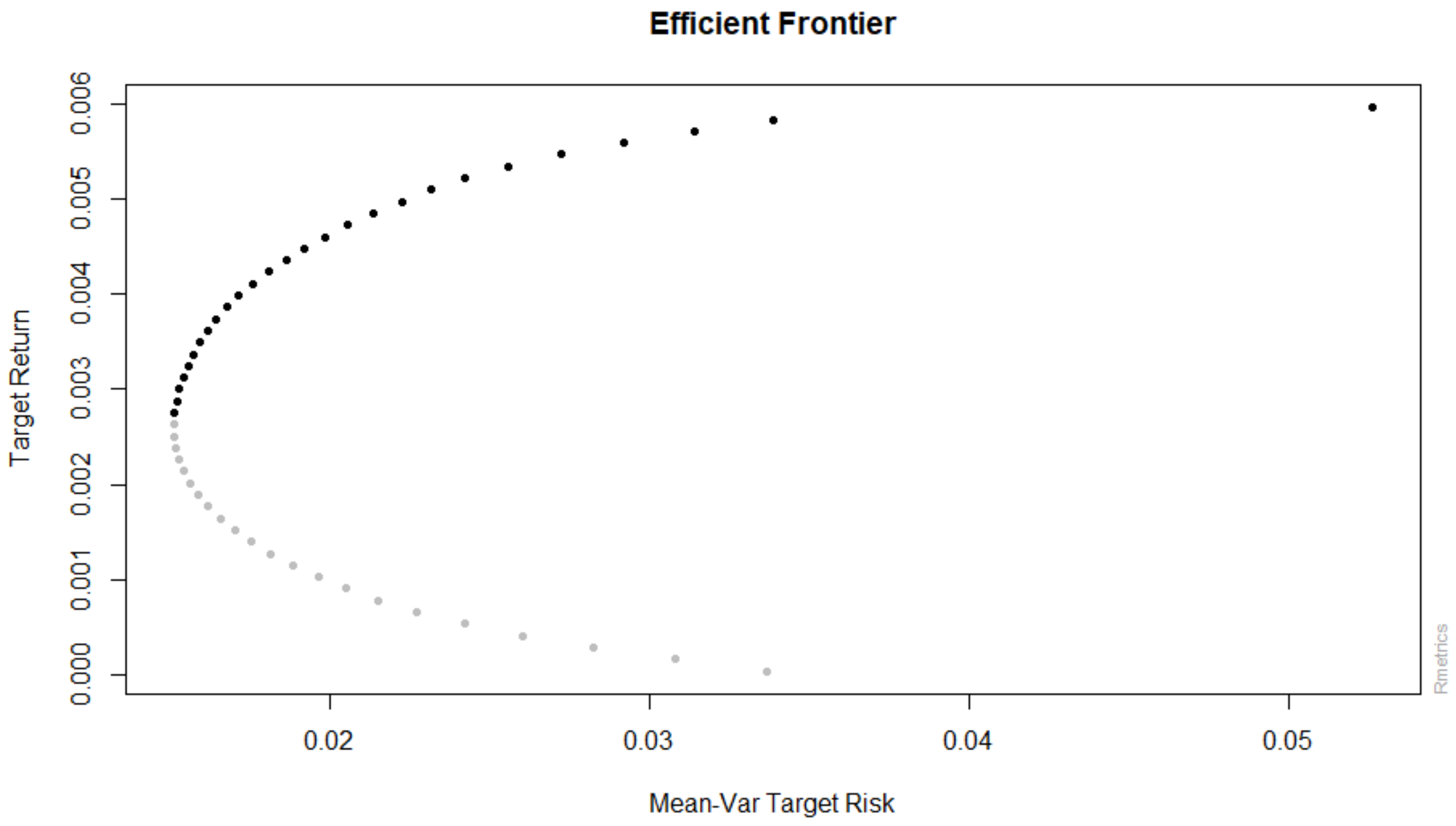
$\Sigma w_i = 1$ → Sum of weights must be 1

$\boldsymbol{w} \geq \mathbf{0}$ → No short selling

Vector of expected returns

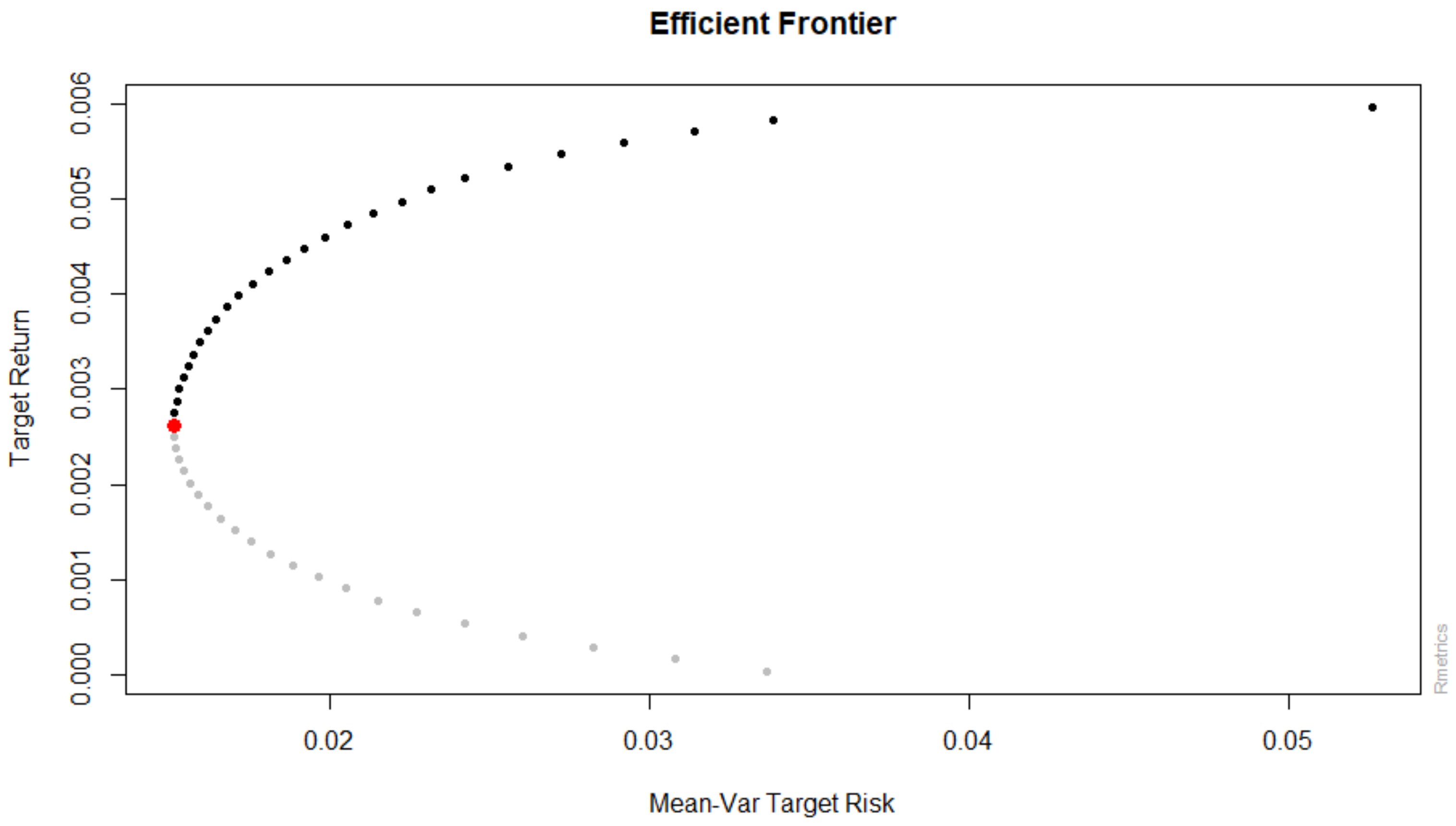
Portfolio Construction

Efficient Frontier



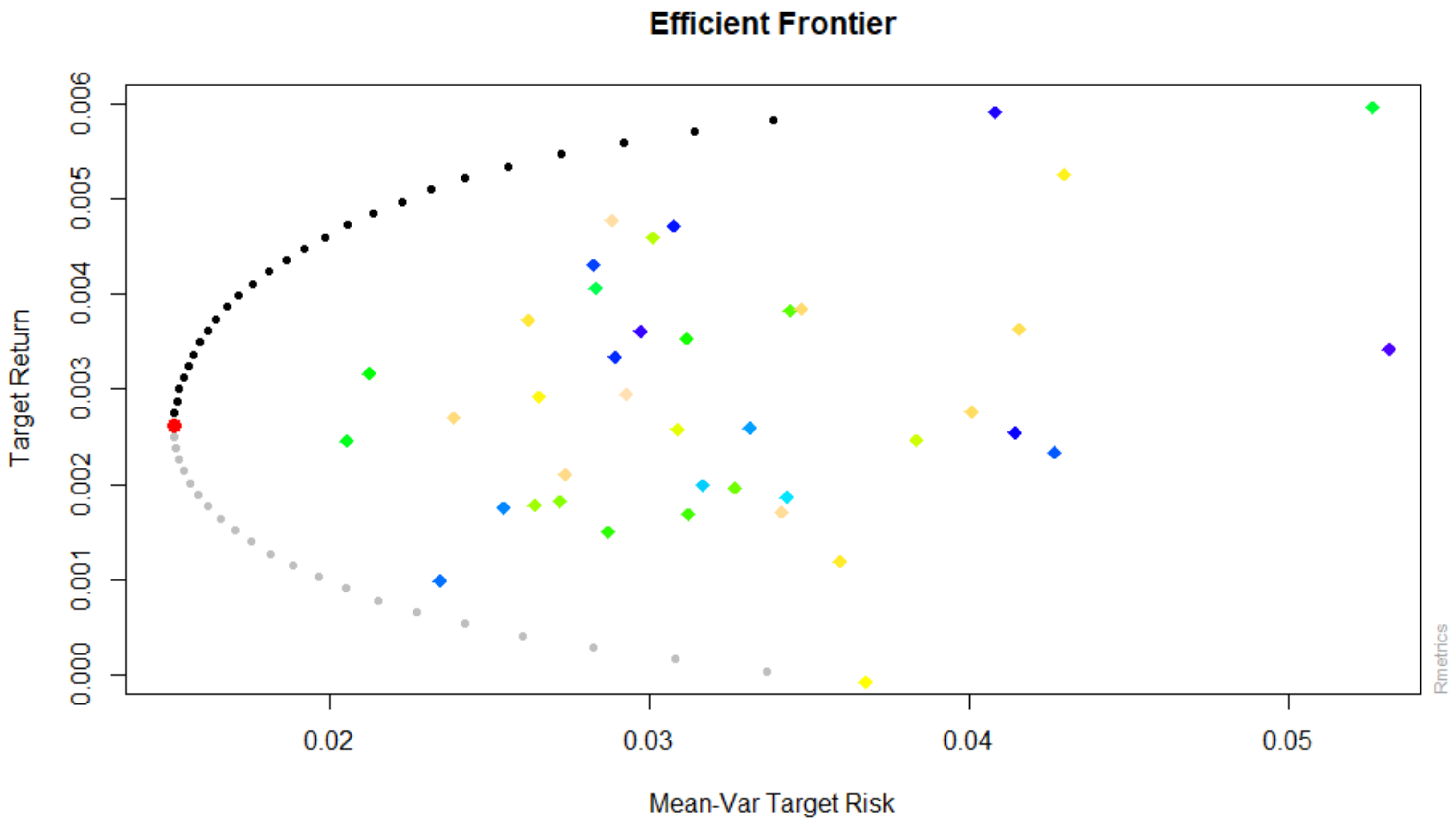
Portfolio Construction

Efficient Frontier with Global Minimum Variance



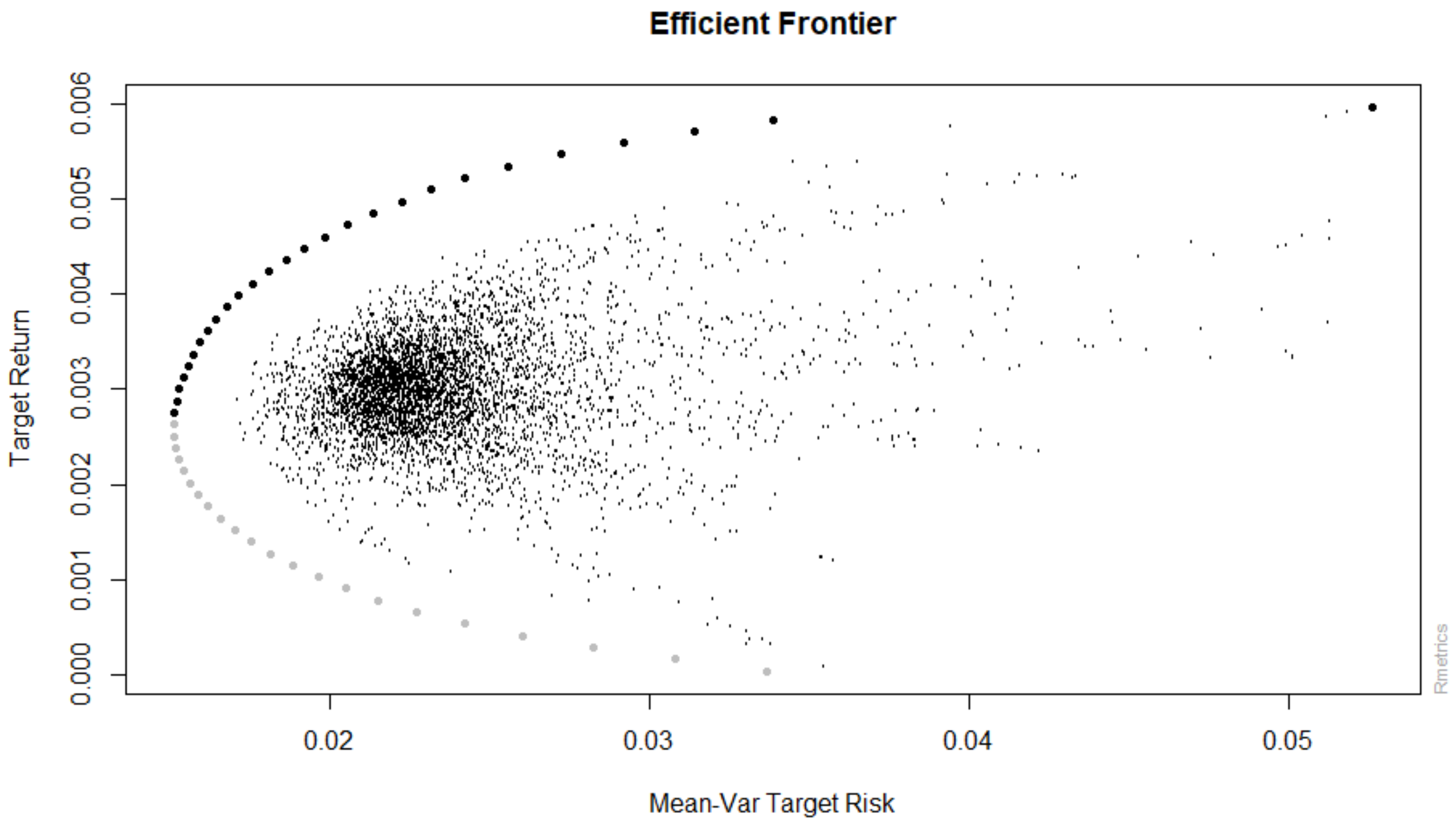
Portfolio Construction

Efficient Frontier with Global Minimum Variance and Risk/Return for Each Asset



Portfolio Construction

Efficient Frontier with Monte Carlo Portfolios



Questions?