Penalised regressions and sparse hedging

Summary

Portfolio choice v2.0

- Classical portfolio theory (mean-variance) is polluted by **estimation errors**
- It can be shown that Minimum Variance weights are closely linked to **cross-sectional regressions**
- In order to reduce estimation risk,[1] it can be useful to resort to *penalised* regressions instead
- The latter can also be used solely for **prediction purposes**

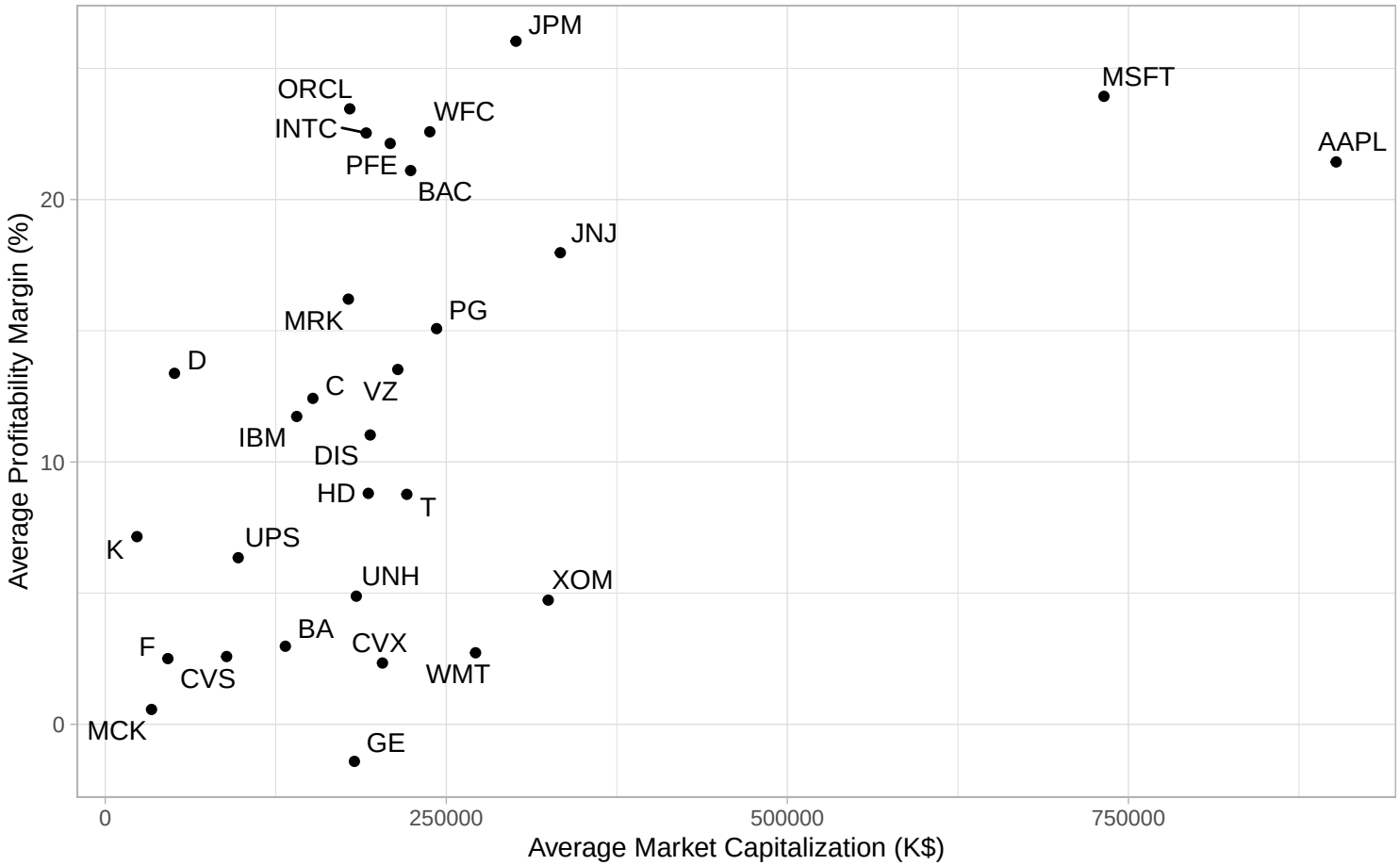
[1] Or fine-tune the bias-variance tradeoff.

Agenda

1. Regular regressions
2. Penalized regressions
3. Bias-variance tradeoff
4. Portfolio considerations
5. Implementation details

For starters: illustrative example (1/4)

Let's say we are given the current data (plotted below)

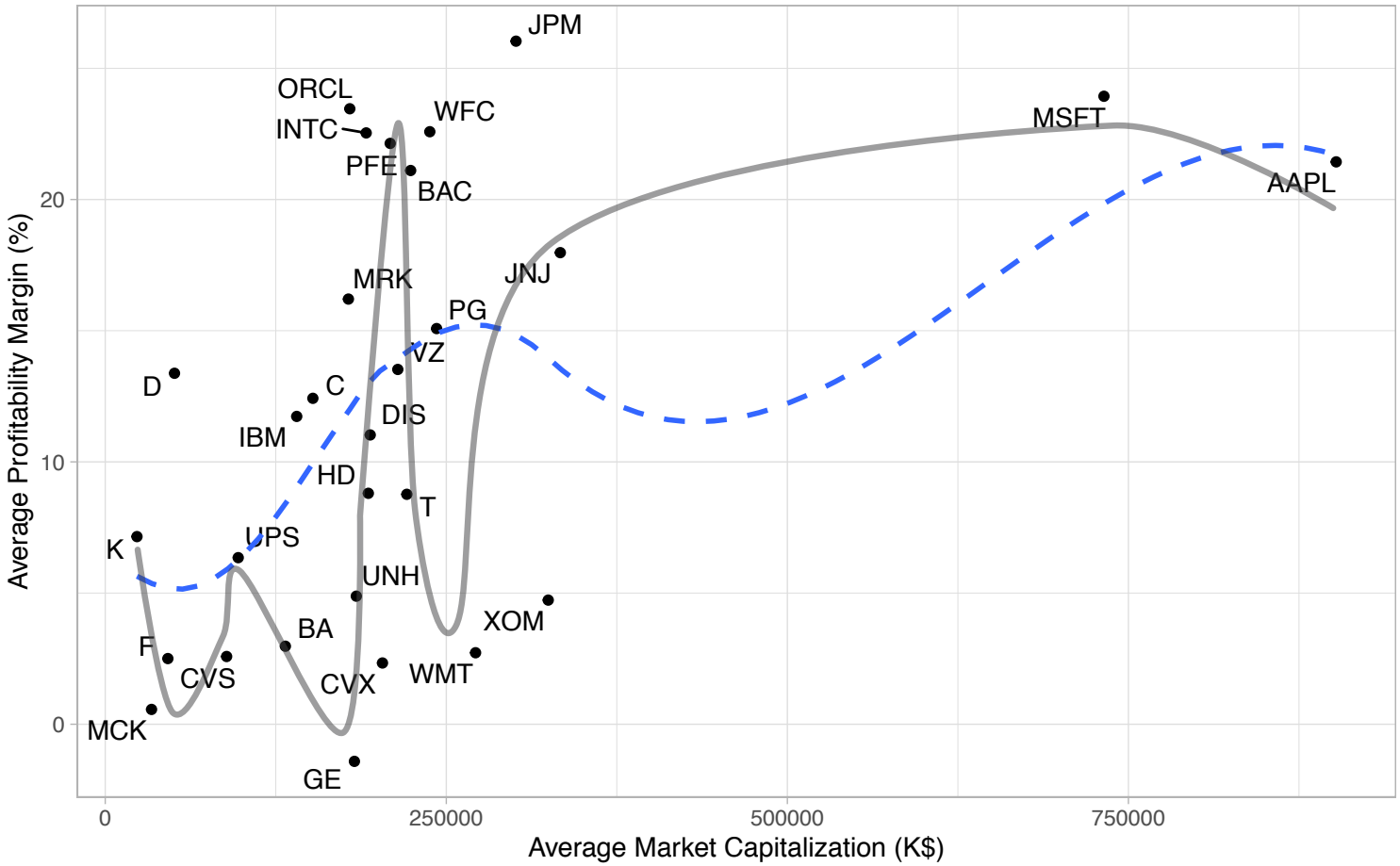


supervised **learning factor** investing

We wish to model the dependence between **profitability** and **size**...

For starters: illustrative example (2/4)

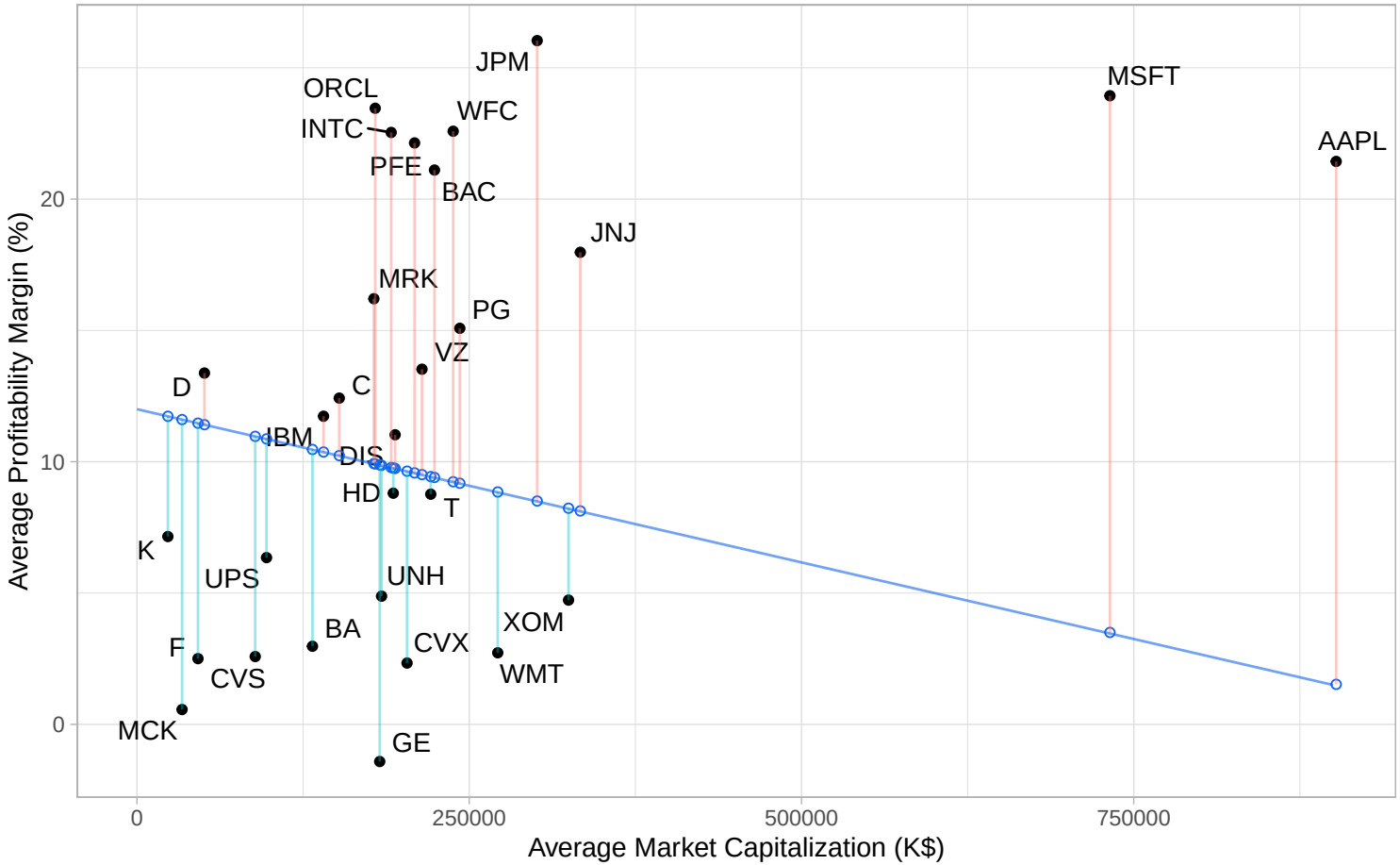
There are tons of ways to find fitting functions:



Maybe simplicity and parsimony are preferable...

For starters: illustrative example (3/4)

The simplest form is the linear function: $prof = a + b * mkt\_cap$.

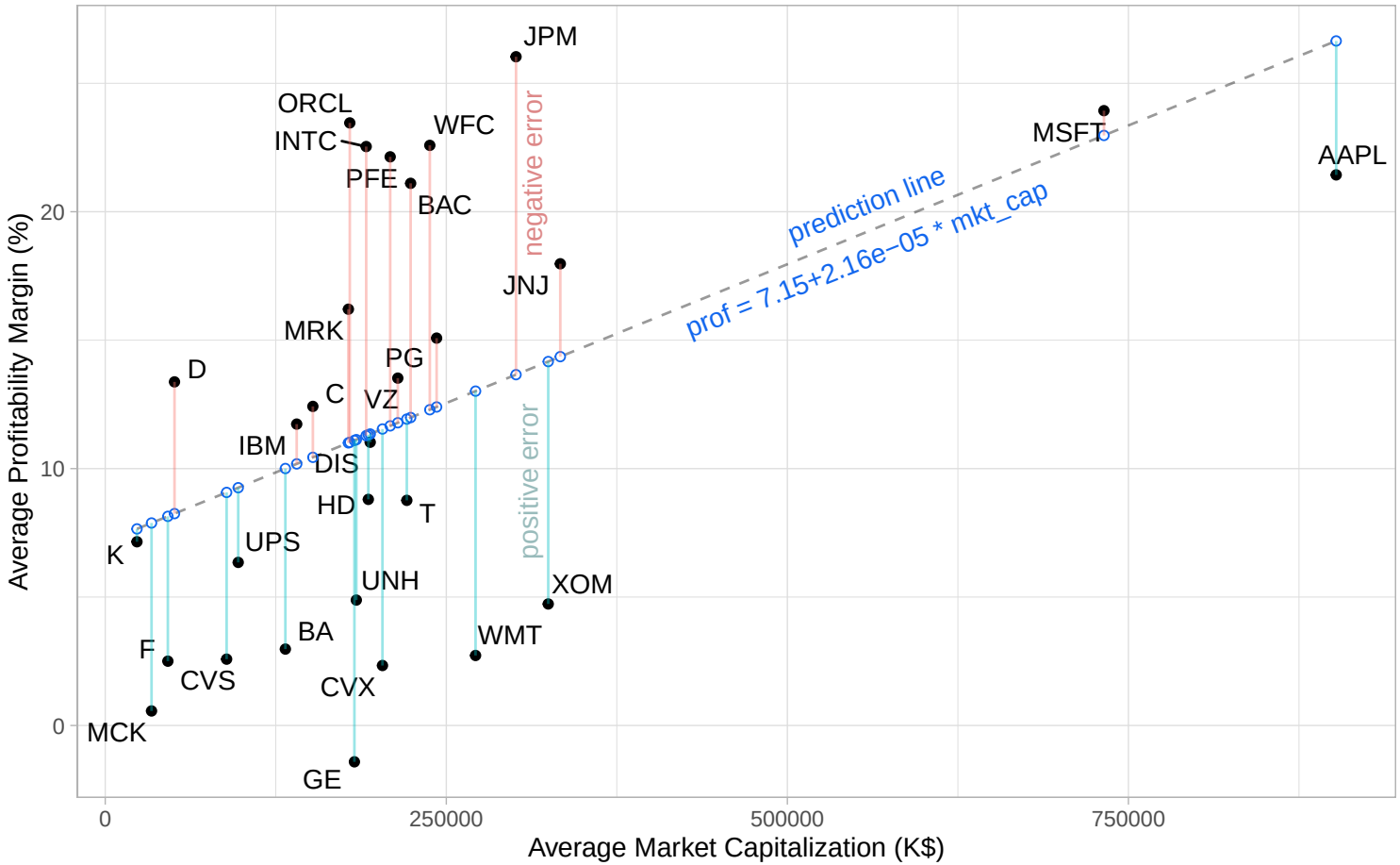


But these errors are huge! You can't choose the line at random!

supervised **learning** **factor** investing

For starters: illustrative example (4/4)

If we minimize the sum of squared errors:



Much better! It's the "*best*" possible line.

Reminder (1/4)

What's a (linear) regression?

You have a **dependent** variable Y and N **exogenous** variables $X^{(n)}$ that are used to 'explain' (or forecast) Y. The model is *linear*, i.e.:

$$Y_t = \beta^{(1)} X_t^{(1)} + \cdots + \beta^{(N)} X_t^{(N)} + \epsilon_t,$$

where t is the index of the observation. If there are T observations, the $\beta^{(n)}$ are usually **estimated** (optimised, in fact) by minimising the quadratic error (sum of squared residuals):

$$\underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{t=1}^{T} \left[Y_t - \left(\beta^{(1)} X_t^{(1)} + \cdots + \beta^{(N)} X_t^{(N)} \right) \right]^2.$$

In compact mode, the model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and the solution to the above problem is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. In terms of dimensions, $\boldsymbol{X}$ is $(T \times N)$, $\boldsymbol{\beta}$ is $(N \times 1)$ so that both $\boldsymbol{Y}$ and $\boldsymbol{\epsilon}$ are T-dimensional (column) vectors.

Reminder (2/4)

Proof of the OLS expression

The sum of squared errors is

$$\begin{aligned}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} &= (\boldsymbol{y} - \boldsymbol{X}\beta)'(\boldsymbol{y} - \boldsymbol{X}\beta) \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\beta'\boldsymbol{X}'\boldsymbol{y} - \beta'\boldsymbol{X}'\boldsymbol{X}\beta\end{aligned}$$

so the gradient (multivariate derivative) is

$$\frac{\partial}{\partial\beta}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = -2\boldsymbol{X}'\boldsymbol{y} - 2\boldsymbol{X}'\boldsymbol{X}\beta$$

and it's equal to $\mathbf{0}$ if and only if $\beta = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$.

Minimization would also require to check the second order condition $\rightarrow -2\boldsymbol{X}'\boldsymbol{X}$ as Hessian matrix.

Reminder (3/4)

Inference vs forecasting

The model is $\boldsymbol{y} = \boldsymbol{X}\beta + e$.

- in **inference**, the aim is to quantify the absolute and relative importance of the variables (columns of $\boldsymbol{X}$) in the way they explain $\boldsymbol{y}$. This is done by making assumptions on the errors $\boldsymbol{e}$, e.g. they follow a Gaussian law and

$$\mathbb{E}[\boldsymbol{e}|\boldsymbol{X}] = \mathbf{0}.$$

- in **forecasting**, the goal is simply to build the best possible model (i.e., that works out of sample). Errors are only interesting if they can be used to improve the accuracy of the model.

Reminder (4/4)

Where problems arise

The usual solution $\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$...

- Obviously, the right part $\boldsymbol{X}'\boldsymbol{Y}$ is not going to be a problem
- The left part $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ on the other hand...
- Whenever the number of features (expl. variables) N is larger than the number of observations (T), the matrix $\boldsymbol{X}'\boldsymbol{X}$ is **singular** (colinearity implies the matrix is not full rank).
- This happens when working with returns when the number of dates is smaller than the number of assets. This explains why daily returns are often required to compute covariance matrices (more on that later).

By the way, a great reference on regression analysis is **Econometric Analysis** by Greene (see Chap. 2→5).

The partitioned form

Another look when a constant is included

The compact form $\boldsymbol{Y} = \alpha \mathbf{1}_T + \boldsymbol{X}\beta + \boldsymbol{\epsilon}$ is solved using the following trick. We define $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \bar{Y}\mathbf{1}_T$ and $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \frac{1}{T}\mathbf{1}_T\mathbf{1}_T'\boldsymbol{X}$, which are the original values, where each column has seen its mean retrieved (i.e., all columns have zero-mean). Then,

$$\hat{\beta} = \left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{Y}},$$

$$\hat{\alpha} = \frac{1}{T}\mathbf{1}_T'(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})$$

where the first term in $\hat{\beta}$ is the **inverse of the covariance matrix** of $\boldsymbol{X}$ and the second term is the vector of **covariances** between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

It is a corollary from the Frisch-Waugh-Lovell theorem (Chap. 3 in Greene).

Bias & Variance

Useful for later..

Assuming a linear **generation** of data: $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$ (where β is unknown), the estimator satisfies

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \epsilon) = \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\epsilon$$

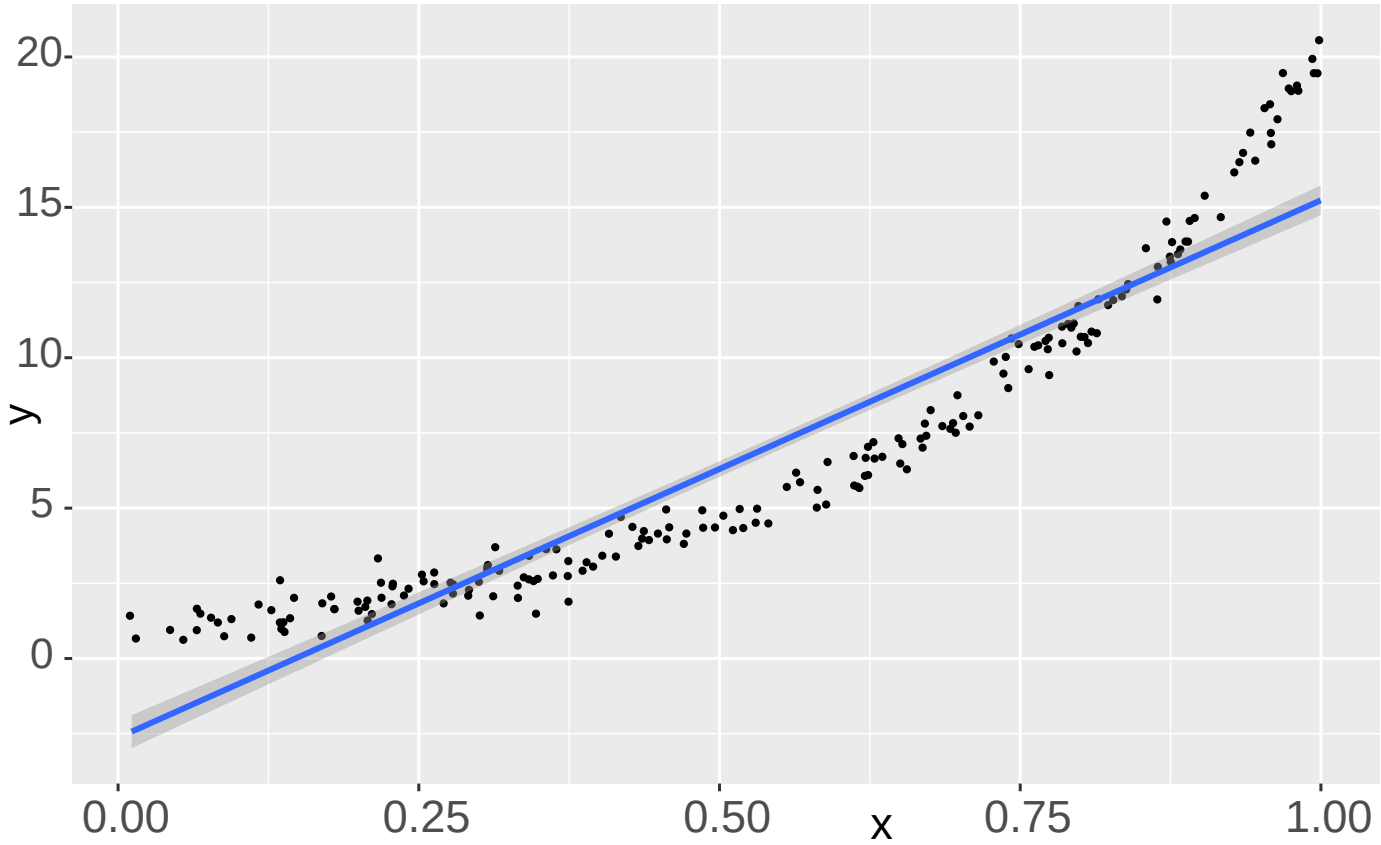
Under the **usual restriction** $\mathbb{E}[\epsilon|\boldsymbol{X}] = 0$, we have $\mathbb{E}[\hat{\beta}] = \beta$. Also,

$$\begin{aligned}\mathbb{V}[\hat{\beta}|\boldsymbol{X}] &= \mathbb{E}[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)|\boldsymbol{X}] \\ &= \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\epsilon\epsilon'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}] \\ &= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1},\end{aligned}$$

as long as $\mathbb{E}[\epsilon\epsilon'] = \sigma^2\boldsymbol{I}_T$.

A key point

About the 'usual' restriction $\mathbb{E}[\epsilon|\boldsymbol{X}] = 0$:



The error terms must have zero mean at any location of X.

Agenda

1. Regular regressions
2. Penalized regressions
3. Bias-variance tradeoff
4. Portfolio considerations
5. Implementation details

Definition(1/2)

Several families of penalised regressions

- **LASSO** (with $\delta > 0$):

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \quad \text{s.t. } ||\boldsymbol{\beta}||_1 \leq \delta$$

- **Ridge** (with $\delta > 0$):

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \quad \text{s.t. } ||\boldsymbol{\beta}||_2^2 \leq \delta$$

- **Elastic-net** (with $\delta > 0$ and $\alpha \in (0, 1)$): linear convex combination of both norms

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \quad \text{s.t. } \alpha||\boldsymbol{\beta}||_1 + (1-\alpha)||\boldsymbol{\beta}||_2^2 \leq \delta$$

We recall the norm notation: $||\beta||_p = \left(\sum_{n=1}^{N} |\beta_n|^p\right)^{1/p}$.

Definition (2/2)

Lagrange form (preferred)

- **LASSO** (with $\lambda > 0$):

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||_1$$

- **Ridge** (with $\lambda > 0$):

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||_2^2$$

- **Elastic-net** (with $\lambda > 0$, $\alpha \in [0, 1]$):

$$\min_{\boldsymbol{\beta}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\alpha ||\boldsymbol{\beta}||_1 + \lambda(1-\alpha)||\boldsymbol{\beta}||_2^2$$

We recall the norm notation: $||\beta||_p = \left(\sum_{n=1}^{N} |\beta_n|^p\right)^{1/p}$.

Ridge: the closed form solution

The only nice form

$$\begin{aligned} L(\beta) &= (\boldsymbol{Y} - \boldsymbol{X}\beta)'(\boldsymbol{Y} - \boldsymbol{X}\beta) + \lambda\beta'\beta \\ &= \boldsymbol{Y}'\boldsymbol{Y} + \beta'\boldsymbol{X}'\boldsymbol{X}\beta - 2\beta'\boldsymbol{X}'\boldsymbol{Y} + \lambda\beta'\beta \end{aligned}$$

so that

$$\frac{\partial L}{\partial \beta} = 2\boldsymbol{X}'\boldsymbol{X}\beta - 2\boldsymbol{X}'\boldsymbol{Y} + 2\lambda\beta$$

and

$$\frac{\partial L}{\partial \beta} = 0 \Leftrightarrow \beta = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_N)^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

where $\boldsymbol{I}_N$ is the $(N \times N)$ identity matrix. The shape is pretty close to the original OLS form. The only (**BIG**) difference is the $\lambda \boldsymbol{I}_N$ inside the matrix inversion. This term ensures that the matrix in indeed **invertible**!

See Ledoit and Wolf (2004) for more details on the shrunk covariance matrix.

Bias & Variance

Slight changes...

Assuming a linear generation of data: $\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon$, the estimator satisfies

$$\begin{aligned}\hat{\beta} &= (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_N)^{-1}\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_N)^{-1}\boldsymbol{X}'(\boldsymbol{X}\beta + \epsilon) \\ &= (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_N)^{-1}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_N - \lambda \boldsymbol{I}_N)\beta + (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_N)^{-1}\boldsymbol{X}'\epsilon\end{aligned}$$

Under the usual restriction $\mathbb{E}[\epsilon|\boldsymbol{X}]$,

$$\mathbb{E}[\hat{\beta}] - \beta = -\lambda(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_N)^{-1}\beta.$$

The bias is not equal to zero. With regard to the variance, it can be shown to be equal to[2]

$$\mathbb{V}[\hat{\beta}] = \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_N)^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_N)^{-1}$$

and goes to zero as λ increases to infinity ($\hat{\beta}$ converges to $\mathbf{0}$).

[2] See Lecture notes on ridge regression by van Wieringen (2018)

Example (0/3)

A simple goal

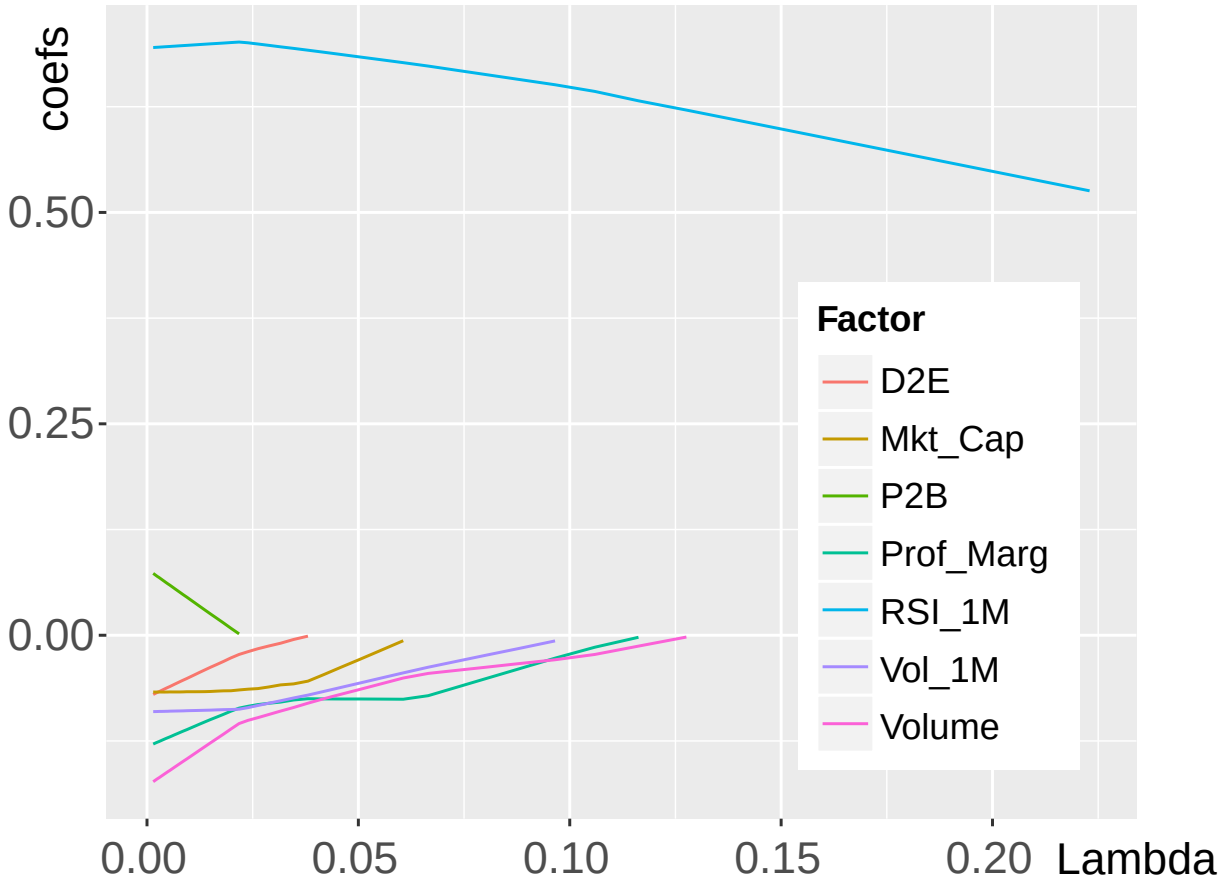
We aim to explain the return of the Apple Inc stock with the help of other **characteristics**:

$$r_t = \alpha + \beta^{Cap} Cap_t + \cdots + \beta^{P2B} P2B_t + \epsilon_t,$$

where the factors (characteristics) have been normalised (because, e.g., Market Cap is out of range).

Example (1/3)

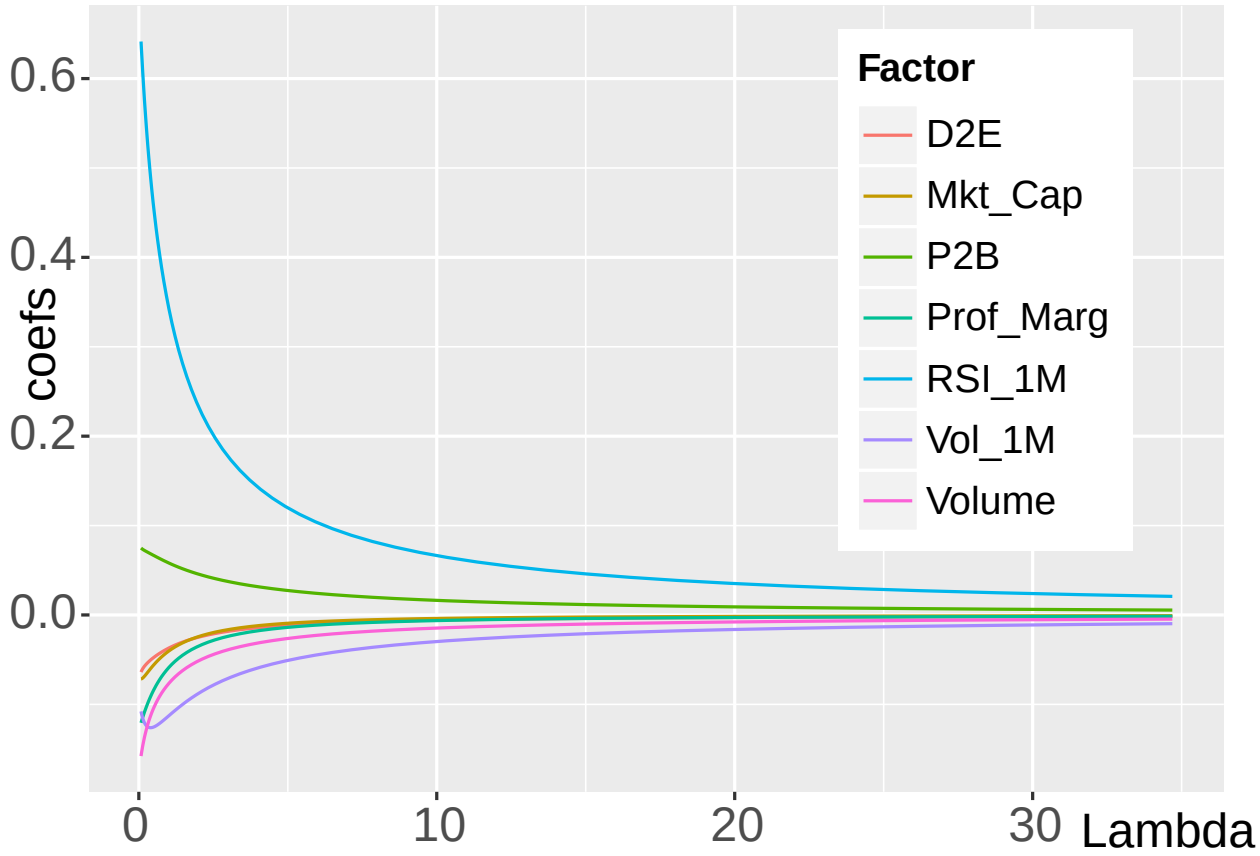
The **Lasso** first:



The coefficients converge to zero pretty quickly!
$\rightarrow$ the Lasso can serve as **feature selection** tool!

Example (2/3)

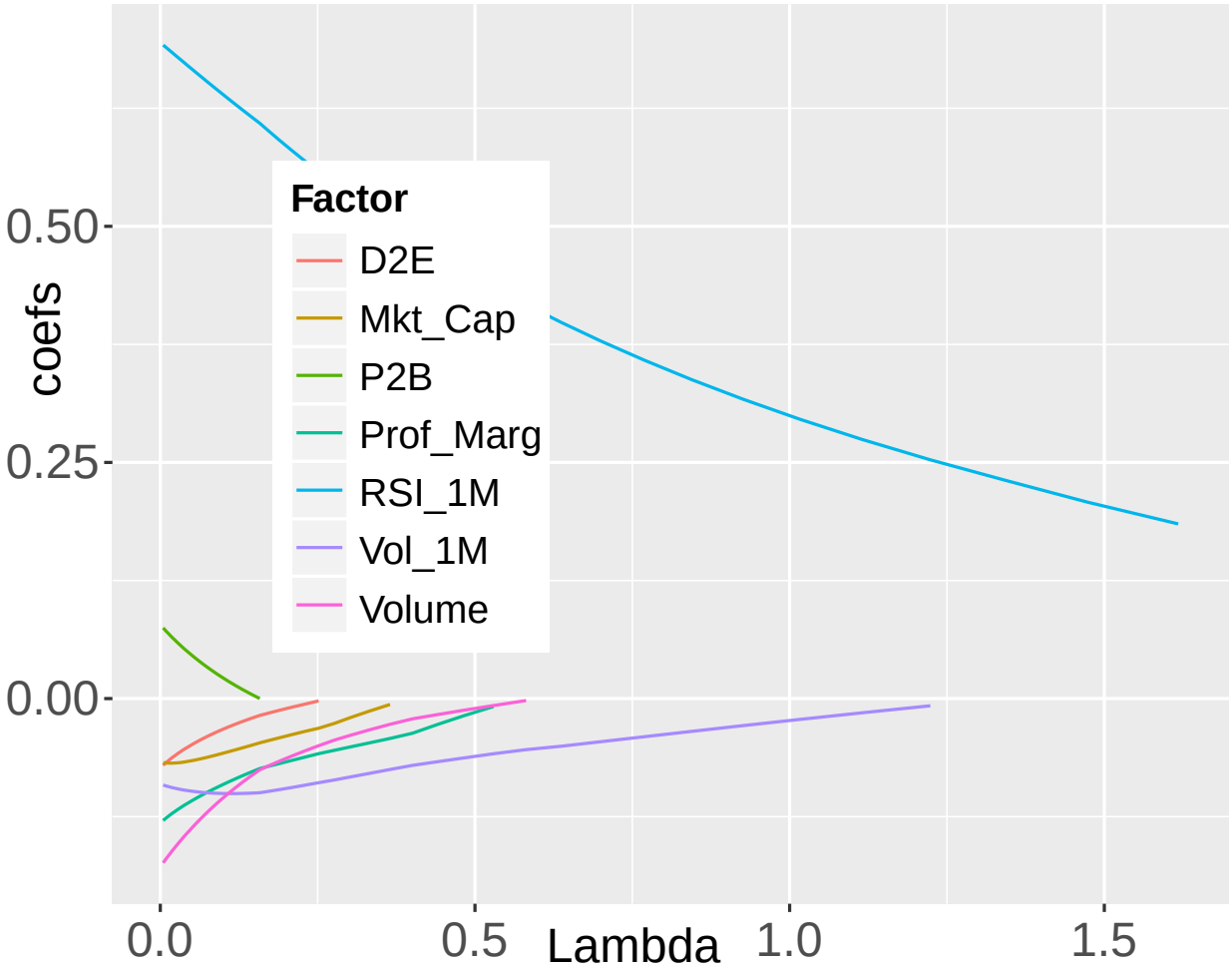
The **Ridge** next:



The coefficients converge to zero very slowly!

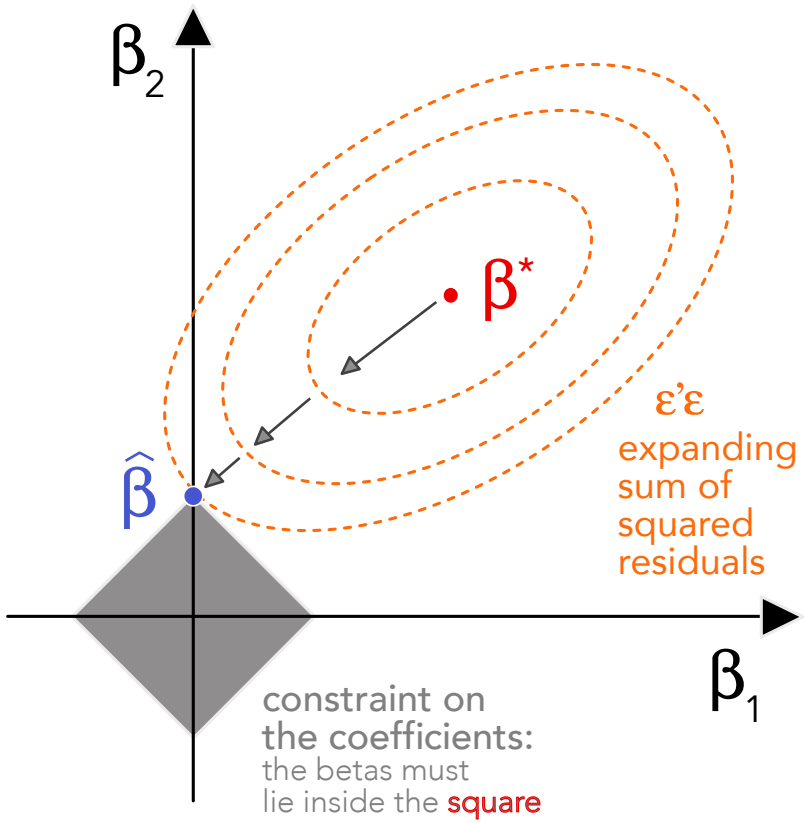
Example (3/3)

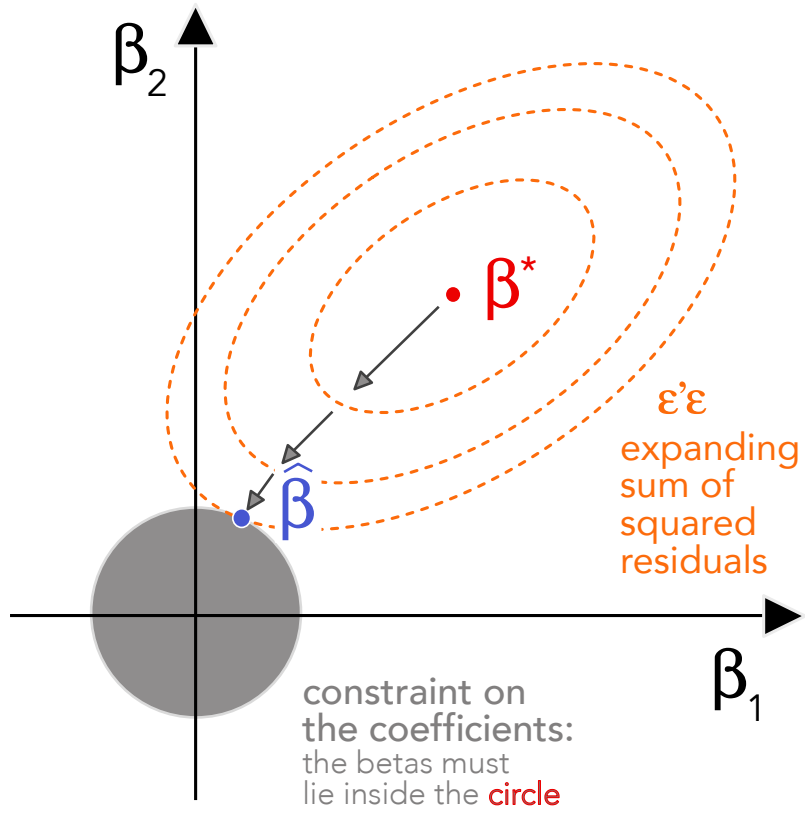
The **elasticnet** at last:



The convergence speed/behaviour is somewhere in the middle!

Why the LASSO is selective





Agenda

1. Regular regressions
2. Penalized regressions
3. Bias-variance tradeoff
4. Portfolio considerations
5. Implementation details

Definition (1/2)

For any model

The aim is to find a model $y = f(x) + \epsilon$ with ϵ having zero mean, σ sd and $\mathbb{E}[\epsilon|x] = 0$. Assume a fitted model $\hat{f}$ (regression, tree, SVM, NN, etc.). The quadratic error at point x is:

$$\begin{aligned}
\mathbb{E}\left[(y - \hat{f}(x))^2\right] &= \mathbb{E}\left[\left(f(x) - \hat{f}(x) + \epsilon\right)^2\right] \\
&= \mathbb{E}\left[\left(f(x) - \hat{f}(x)\right)^2\right] + \sigma^2 \\
&= \underbrace{\left(\mathbb{E}\left[f(x) - \hat{f}(x)\right]\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}\left[\hat{f}(x)^2\right] - \mathbb{E}\left[\hat{f}(x)\right]^2}_{\text{Variance}} + \sigma^2
\end{aligned}$$

In the above expression, the *sample* x is **random**.

Definition (2/2)

In the above expression:

- the **bias** assesses the *average* distance between the *true* f and the fitted one $\hat{f}$,
- the **variance** measures the dispersion of $\hat{f}$ over all possible x,
- σ is the **irreducible error**: it is the level of approximation reached by the true f.

A very complex model is going to reduce bias but by doing so, it will move a lot to capture a maximum of points. To the other end of the spectrum, a constant model $\hat{f}(x) = c$ has zero variance, but a rather large bias.

Agenda

1. Regular regressions
2. Penalized regressions
3. Bias-variance tradeoff
4. Portfolio considerations
5. Implementation details

References

The interested reader should have a look at:

- **On the Inverse of the Covariance Matrix in Portfolio Analysis** (Stevens, JF 1998)
- **Improving Mean Variance Optimization through Sparse Hedging Restrictions** (Goto and Xu JFQA 2015)

See also:
The Sampling Error in Estimates of Mean-Variance Efficient Portfolio Weights (Britten-Jones, JF 1999)

Mean-variance portfolios

The textbook derivation (no budget constraint: simplicity)

Minimise variance for a fixed return... and hope for the best!

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w}, \quad \text{s.t. } \boldsymbol{w}'\boldsymbol{\mu} = r^*$$

with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ first and second moment of expected (excess) returns.

$$L(\boldsymbol{w}) = \boldsymbol{w}'\boldsymbol{\Sigma}\boldsymbol{w} - \lambda(\boldsymbol{w}'\boldsymbol{\mu} - r^*)$$

Hence

$$\frac{\partial L}{\partial \boldsymbol{w}} = 2\boldsymbol{\Sigma}\boldsymbol{w} - \lambda\boldsymbol{\mu} = 0 \Leftrightarrow \boldsymbol{w} = \frac{\lambda}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},$$

where the λ scaling factor relates to risk preferences.

In practice, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are **estimated**. Usually, $\boldsymbol{\mu}$ is the one that is hardest to evaluate (and has the most importance!).

MV portfolios as regression estimates!

A surprising identiy

Britten-Jones (1999) remarked that you can interpret:

$$\boldsymbol{w} = \underbrace{\Sigma^{-1}}_{(\boldsymbol{X}'\boldsymbol{X})^{-1}} \times \underbrace{\mu}_{\boldsymbol{X}'\mathbf{1}}$$

Thus, if $\boldsymbol{X}$ is a matrix of returns, $\boldsymbol{w}$ is the OLS estimate of:

$$\mathbf{1} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{e}$$

Amazing!

Minimum variance portfolios

The textbook derivation (with budget constraint here)

Minimise variance and that's all!

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\Sigma\boldsymbol{w}, \quad \text{s.t. } \boldsymbol{w}'\mathbf{1}_N = 1,$$

where the last constraint is simply the **budget constraint**. Then,

$$L(\boldsymbol{w}) = \boldsymbol{w}'\Sigma\boldsymbol{w} - \lambda(\boldsymbol{w}'\mathbf{1}_N - 1)$$

Hence

$$\frac{\partial L}{\partial \boldsymbol{w}} = 2\Sigma\boldsymbol{w} - \lambda\mathbf{1}_N = 0 \Leftrightarrow \boldsymbol{w} = \frac{\lambda}{2}\Sigma^{-1}\mathbf{1}_N,$$

where the λ is determined by the budget:

$$\boldsymbol{w} = \frac{\Sigma^{-1}\mathbf{1}_N}{\mathbf{1}_N'\Sigma^{-1}\mathbf{1}_N}.$$

IN ANY CASE, Σ^{-1} **IS REQUIRED**!

A deep interpretation (1/3)

A closer look at the inverse

If we **decompose** the matrix Σ into:

$$\Sigma = \begin{bmatrix} \sigma^2 & \boldsymbol{c}' \\ \boldsymbol{c} & \boldsymbol{C} \end{bmatrix},$$

classical **partitioning** results (e.g., via Schur complements) imply

$$\Sigma^{-1} = \begin{bmatrix} (\sigma^2 - \boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{c})^{-1} & -(\sigma^2 - \boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{c})^{-1}\boldsymbol{c}'\boldsymbol{C}^{-1} \\ -(\sigma^2 - \boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{c})^{-1}\boldsymbol{C}^{-1}\boldsymbol{c} & \boldsymbol{C}^{-1} + (\sigma^2 - \boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{c})^{-1}\boldsymbol{C}^{-1}\boldsymbol{c}\boldsymbol{c}'\boldsymbol{C}^{-1} \end{bmatrix}.$$

We are interested in the first line, which has 2 components: the factor $(\sigma^2 - \boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{c})^{-1}$ and the line vector $\boldsymbol{c}'\boldsymbol{C}^{-1}$. $\boldsymbol{C}$ is the covariance matrix of assets 2 to N and $\boldsymbol{c}$ is the covariance between the first asset and all other assets. The first line of Σ^{-1} is

$$(\sigma^2 - \boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{c})^{-1} \begin{bmatrix} 1 & \underbrace{-\boldsymbol{c}'\boldsymbol{C}^{-1}}_{N-1 \text{ terms}} \end{bmatrix}.$$

A deep interpretation (2/3)

A special regression

We regress the returns of the first asset against those of all other assets:

$$r_{1,t} = a_1 + \sum_{n=2}^{N} \beta_{1|n} r_{n,t} + \epsilon_t, \text{ i.e., } \boldsymbol{r}_1 = a_1 \mathbf{1}_T + \boldsymbol{R}_{-1}\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$$

The **OLS estimator** for $\boldsymbol{\beta}_1$ is

$$\hat{\boldsymbol{\beta}}_1 = \boldsymbol{C}^{-1}\boldsymbol{c}$$

and in addition (exercise!):

$$(1 - R^2)\sigma^2_{\boldsymbol{r}_1} = \sigma^2_{\boldsymbol{r}_1} - \boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{c} = \sigma^2_{\epsilon,1}$$

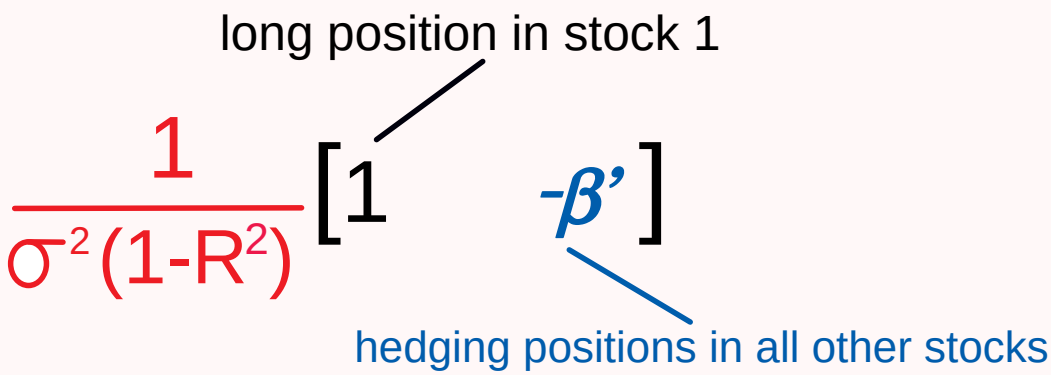
(where in fact $\sigma^2_{\boldsymbol{r}_1} = \sigma^2$) and the first line of $\boldsymbol{\Sigma}^{-1}$ is equal to

$$\frac{1}{\sigma^2_{\epsilon,1}} \times \begin{bmatrix} 1 & -\hat{\boldsymbol{\beta}}_1' \end{bmatrix}$$

A deep interpretation (3/3)

Remember Minimum Variance?

Each line of the inverse covariance matrix is multiplied by $\mathbf{1}_N$, and one line has form:



The red scaling factor is the inverse of $\hat{\epsilon}'\hat{\epsilon}$: when errors are large, positions are smaller. Note: β was chosen to **minimise the error/variance** pertaining to the above portfolio, which is a clear reference to 'minimum variance'.

This applies to all other lines/stocks!

Bottom line

Reduce estimation risk

- When T is not much larger than N, the **covariance matrix** is noisy, and its inverse even more. It is possible to **regularise** the estimation of these matrices upfront (see the work of Ledoit and Wolf).
- Nonetheless, the regularisation can occur at the stock level, using **hedging relationships**.
- The generalisation is straightforward: **bias** vs **tradeoff** AND **transaction costs** imply that penalised regression *could* be a good idea.

Agenda

1. Regular regressions
2. Penalized regressions
3. Bias-variance tradeoff
4. Portfolio considerations
5. Implementation details

The strategies (1/2)

Sparse portfolios.

Pseudo-algorithm

For all dates t,

- For all stocks i,
  1. estimate the elastic-net regression over the $t = 1, \ldots, T$ samples

$$\underset{\beta_{i|}}{\text{argmin}} \left\{ \sum_{t=1}^{T} \left(r_{i,t} - a_i - \sum_{n \neq i}^{N} \beta_{i|n} r_{n,t} \right)^2 + \lambda\alpha ||\beta_{i|}||_1 + \lambda(1-\alpha)||\beta_{i|}||_2^2 \right\}$$

  2. to get the weights of asset i: $w_i = \frac{1 - \sum_{j \neq i} \beta_{i|j}}{\sigma^2_{\epsilon,i}}$

Note, when the number of asset is large, this may take too much time and the regularisation can be performed directly on the inverse covariance matrix (for Lasso and Ridge at least): precision matrix estimation via the **GLASSO** for instance.

The strategies (2/2)

Predictive LASSO

Pseudo-algorithm

For all dates t,

- For all stocks i,
  1. estimate a predictive model

$$r_{n,t} = \alpha + \sum_{k=1}^{K} \beta_n^k f_{t-1}^k + \epsilon_{n,t}, \quad s.t. \quad \alpha||\boldsymbol{\beta}_n||_1 + (1-\alpha)||\boldsymbol{\beta}_n||_2^2 \leq \delta$$

  2. form a prediction with current factor/feature values f_t^k
  3. devise an allocation scheme based on this prediction

There is a large palette of potential **predictors** f_t^k (e.g., 'true' asset pricing factors), but we will simply work with (rescaled) firm attributes.

Thank you for your attention

Any questions?

Proof for R^2

With $\boldsymbol{X}$ being the concatenation of $\mathbf{1}_T$ with returns $\boldsymbol{R}_{-1}$ and with $\boldsymbol{y} = \boldsymbol{r}_1$, $R^2 = 1 - \frac{\boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{T\sigma_Y^2} = 1 - \frac{\boldsymbol{y}'\boldsymbol{y} - \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}}{T\sigma_Y^2} = 1 - \frac{\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\hat{\boldsymbol{\beta}}}{T\sigma_Y^2}$, with fitted values $\boldsymbol{X}\hat{\boldsymbol{\beta}} = \hat{a}_1\mathbf{1}_T + \boldsymbol{R}_{-1}\boldsymbol{C}^{-1}\boldsymbol{c}$. Hence,

$$
\begin{aligned}
T\sigma_{\boldsymbol{r}_1}^2 R^2 &= T\sigma_{\boldsymbol{r}_1}^2 - \boldsymbol{r}_1'\boldsymbol{r}_1 + \hat{a}_1\mathbf{1}_T'\boldsymbol{r}_1 + \boldsymbol{r}_1'\boldsymbol{R}_{-1}\boldsymbol{C}^{-1}\boldsymbol{c} \\
T(1-R^2)\sigma_{\boldsymbol{r}_1}^2 &= \boldsymbol{r}_1'\boldsymbol{r}_1 - \hat{a}_1\mathbf{1}_T'\boldsymbol{r}_1 - \left(\tilde{\boldsymbol{r}}_1 + \frac{\mathbf{1}_T\mathbf{1}_T'}{T}\boldsymbol{r}_1\right)'\left(\tilde{\boldsymbol{R}}_{-1} + \frac{\mathbf{1}_T\mathbf{1}_T'}{T}\boldsymbol{R}_{-1}\right)\boldsymbol{C}^{-1}\boldsymbol{c} \\
T(1-R^2)\sigma_{\boldsymbol{r}_1}^2 &= \boldsymbol{r}_1'\boldsymbol{r}_1 - \hat{a}_1\mathbf{1}_T'\boldsymbol{r}_1 - T\boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{c} - \boldsymbol{r}_1'\frac{\mathbf{1}_T\mathbf{1}_T'}{T}\boldsymbol{R}_{-1}\boldsymbol{C}^{-1}\boldsymbol{c} \\
T(1-R^2)\sigma_{\boldsymbol{r}_1}^2 &= \boldsymbol{r}_1'\boldsymbol{r}_1 - \frac{(\mathbf{1}_T'\boldsymbol{r}_1)^2}{T} - T\boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{c} \\
(1-R^2)\sigma_{\boldsymbol{r}_1}^2 &= \sigma_{\boldsymbol{r}_1}^2 - \boldsymbol{c}'\boldsymbol{C}^{-1}\boldsymbol{c}
\end{aligned}
$$

where in the fourth equality we have plugged $\hat{a}_1 = \frac{\mathbf{1}_T'}{T}(\boldsymbol{r}_1 - \boldsymbol{R}_{-1}\boldsymbol{C}^{-1}\boldsymbol{c})$.
(there is probably a simpler proof - see e.g. Section 3.5 in Greene)