Big Data in Asset Management[1]

Thierry Roncalli[†‡]

[†]Professor of Economics, Évry University, France

[‡]Head of Research & Development, Lyxor Asset Management, France

ESMA/CEMA/GEA meeting, CNMV, Madrid, Spain

November 13, 2014

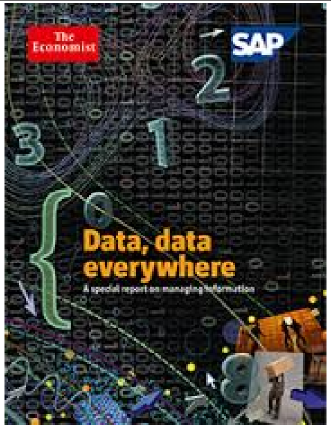
[1]The opinions expressed in this presentation are those of the author and are not meant to represent the opinions or official positions of Lyxor Asset Management. I would like to thank Antoine Frachot, Head of GENES, Kamel Gadouche, Head of CASD, Sébastien Roussel and Alain Viénot from CASD for their helpful comments and for providing the materials on the CASD technology. I also thank David Bessis from TINYCLUES for providing some charts used in this presentation.

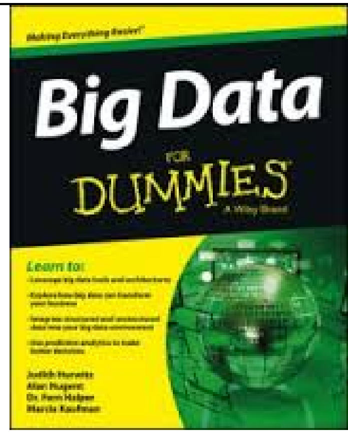
Outline

What Does Big Data Mean?

- Definition
- Analysis of Big Data Problems
- The Missing Factor

A very hot topic





- Nature, Science, The Economist, MacKinsey.
- Obama, White House, etc.
- New York Times (The Age of Big Data), Wall Street Journal (Meet the New Boss: Big Data), Financial Times (Acute 'Big Data' Challenges Facing Asset Managers), etc.
- Davos World Economic Forum (Big Data, Big Impact): Big data = new asset class.
- Le Monde, Libération, Le Figaro, l'Agefi, 20 Minutes, France Inter, etc.
- Canal+ (Big Data : les nouveaux devins).
- Rapport Lauvergeon 'Innovation 2030'

Definition

McKinsey Global Institute (2011)

Big data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze.

$\Rightarrow$ There are certainly few big data problems.

Various aspects

- Large dataset (Megabyte, Gigabyte, Terabyte, Petabyte, Exabyte)
- Unstructured data (networked data but fuzzy relationships)
- Data-driven research, business & decisions
- High skills (IT, statistics, etc.)

$\Rightarrow$ Big data problems differ from one sector to another.

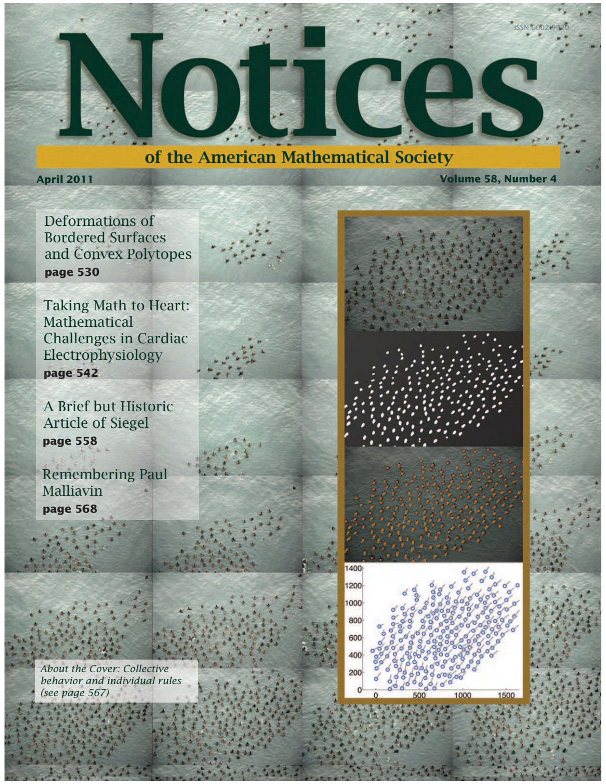
$\boxed{\text{Big data} \neq \text{unified science}}$

Application domains of big data

- Web analytics
- Pattern recognition
- Personal location tracking
- Text analysis
- Public sector administration (government)
- Health care
- Scientific research (human genome, etc.)
- Targeted marketing
- Retail/customer behavior
- Credit scoring
- Banking transactions
- Finance

Big data blurs the frontiers between sciences

What is the link between mathematics and biology?



Big data blurs the frontiers between sciences

The answer is finance:

QUANTITATIVE FINANCE

RENAISSANCE TECHNOLOGIES, a quantitatively based financial management firm, has openings for research and programming positions at its Long Island, NY research center.

Research & Programming Opportunities

We are looking for highly trained professionals who are interested in applying advanced methods to the modeling of global financial markets. You would be joining a group of roughly one hundred fifty people, half of whom have Ph.D.s in scientific disciplines. We have a spectrum of opportunities for individuals with the right scientific and computing skills. Experience in finance is not required.

The ideal research candidate will have:

- **A Ph.D. in Computer Science, Mathematics, Physics, Statistics, or a related discipline**
- **A demonstrated capacity to do first-class research**
- **Computer programming skills**
- **An intense interest in applying quantitative analysis to solve difficult problems**

The ideal programming candidate will have:

- **Strong analytical and programming skills**
- **An In depth knowledge of software development in a C++ Unix environment**

Compensation is comprised of a base salary and a bonus tied to company-wide performance.

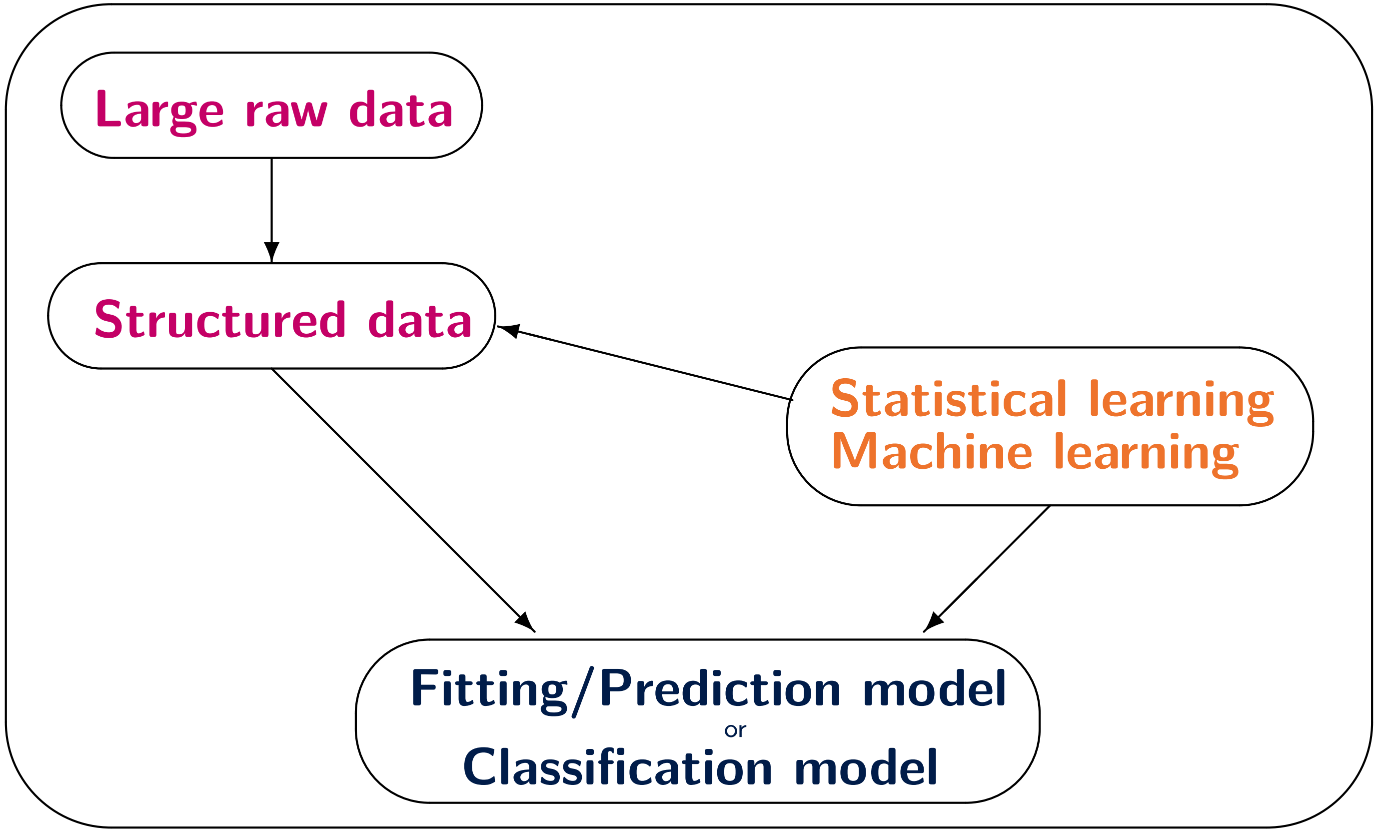
Send a *copy of your resume to: careers@rentec.com*
No telephone inquiries.

Renaissance

An equal opportunity employer.

Architecture of big data problems

Figure: Supervised / unsupervised learning problems



The big challenge (1/2)

Gartner's 3V model of big data:
Volume (amount of data)
Velocity (speed of data)
Variety (data types)

$\Downarrow$

4V model:
+ **V**eracity/**V**alue

The challenge

How to transform raw data into structured (informative) data?

1. Heterogeneous variables $\Rightarrow$ comparable and workable data
2. Heterogeneous variables $\Rightarrow$ new variables
3. Heterogeneous variables $\Rightarrow$ valuable variables
4. Heterogeneous variables $\Rightarrow$ model variables

The big challenge (1/2)

Raw data | **Model variables** | **Prediction**

$\mathbf{X}$ | $\mathbf{Y} = \mathbf{f}(\mathbf{X})$ | $\mathbf{Z} = \mathbf{g}(\mathbf{Y})$

The most difficult step is transforming X into Y:

- Averaging, Averaging^2, Averaging^3, etc.
- Cutting X_1 into Y_1, Y_2, etc.
- Aggregating X_1, X_2, etc. into Y_1
- Creating classes from X_1, X_2, etc.
- Conditioning X_1 by X_2, etc.
- Dummy variables everywhere!

$\Rightarrow$ **The Y variables are more important than the model g itself!**

The missing factor

The missing factor is the construction of the database:

- Structure of databases are generally difficult to change;
- Most of the time, data are located in several databases;
- Missing items have a big impact;
- Etc.

⇒ It is not a big data issue, but a challenge for the information system (IS).

$$\boxed{\textbf{BIG DATA} = \textbf{DATA} + \ldots}$$

Responsibilities and data management

Goal

Building a comprehensive database of the hedge fund industry.

One solution is to merge existing commercial databases (HFR, EurekaHedge, BarclayHedge, Morningstar, Lipper TASS, etc).

$\Rightarrow$ Not simple.

How to complete this database with Newcits funds?

Who manages the project?

- IT?
- Experts?

The chicken-and-egg problem

An illustration

- January 2001: Second Consultative Paper for a New Basel Capital Accord
- 2001: A race for building operational risk loss databases (internal & external)
- Frachot and Roncalli (2002, 2003) document reporting biases in loss data and show that the computation of the value-at-risk needs data collection thresholds.
- Big impact on the design of operational risk loss databases

What is the puzzle?

- You need data in order to test the model (data $\Rightarrow$ model)
- You need to test the model before designing the database (model $\Rightarrow$ data)

What is a data Scientist?

- 1970-1990: data are managed by statisticians (light projects)
- 1990-2010: data are managed by IT people (heavy projects)

$\Rightarrow$ Now, data are managed by data scientists (computer science, modeling, statistics and analytics).

Why?

- Quick-and-dirty $\rightarrow$ slow-and-robust
- "Data first, then model" is **WRONG**
- Don't get lost in the IS
- Liability of the project

Relationships between the industry and academics

Fiction

Consider an asset management company who has a comprehensive dataset of order books (European markets, stocks, futures, ETFs, options, etc.).

This asset manager would like to sponsor a research chair and give academics access to the database in order to study the ETF industry (liquidity, volatility, micro-structure feedback, systemic risks, etc.).

How to proceed?

1. FTP?
2. Internships?
3. Academic consultants?
4. Other solutions?

Objective

Value of the data

- Data means information
- Data may have a cost
- Data may be sensitive or strategic
- Data may be confidential

$\Rightarrow$ Data, data everywhere, but data are poorly exploited.

We prefer that data are not used rather than they get out of our control.

The objective is then to control data access, maximize outputs and minimize dissemination risk.

The CASD technology

Remark

The following slides concerning CASD have been kindly provided by the CASD team.

For more information:

- *Kamel Gadouche, Director of CASD, kamel.gadouche@casd.eu*
- *Antoine Frachot, Head of GENES, antoine.frachot@groupe-genes.fr*

The CASD technology

Groupe des Ecoles Nationales d'Economie et Statistique (GENES)

- An institution of higher education and research in the field of economics, finance, etc (ENSAE, ENSAI, CREST, CEPE).
- Under the umbrella of the French Ministry of the Economy
- The following governmental directorates are represented in the GENES Board: French Treasury, Bank of France, INSEE, Ministry of Finance, Ministry of Industry, Ministry of Higher Education.

2007-2008: GENES started developing a remote safe center for confidential administrative data (Centre d'Accès Securisé aux Données or CASD).

$\Rightarrow$ CASD is the official platform which hosts 'sovereign' (and therefore confidential) data coming from the French administration and dedicated to research in economics, social sciences and public policies.

The CASD technology

The SD Box

The CASD technology

How does it work?

Dedicated to securely accessing sensitive data

Rigorous & Imperative Authentication

Physically & Digitally Locked to limit file retrievals

No sensitive data stored in SD-Box

Standardized Units

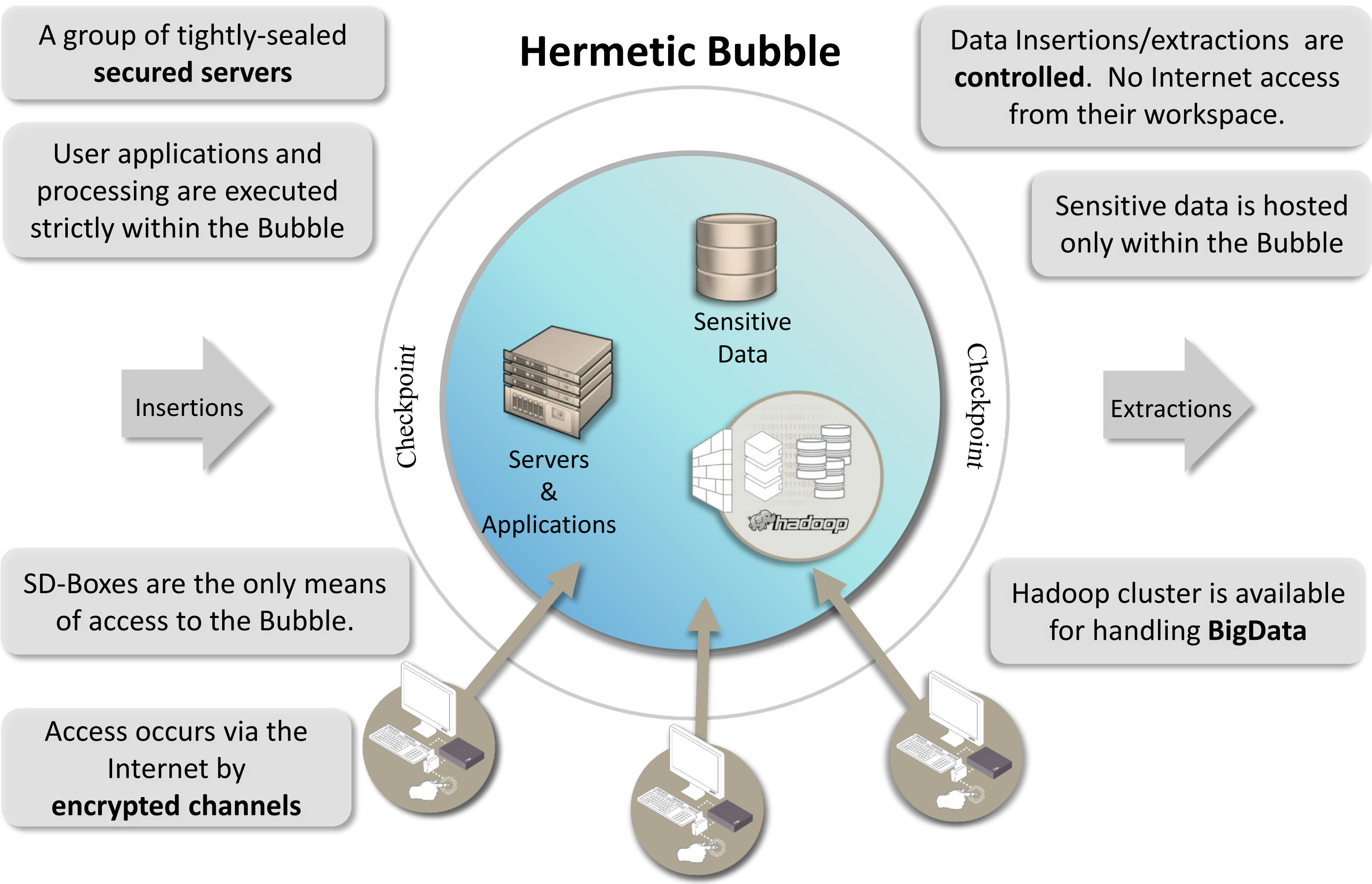
- Reliability (Unique, Stable & Tested Configuration)
- Simple & Economical Set-Up
- Very Limited Need for Assistance

Easy Installation

- Screen, Keyboard & Mouse Needed
- Internet Connection Required
- Simple to Configurate
- No Effects on the Rest of the IS

The CASD technology

The infrastructure



The CASD technology

For what purpose?

The objective

The Secure Data Access Centre (CASD) has been designed to address the issues of sensitive data dissemination.

- Confidential data / Official data
- Collaboration with CNIL[a] to address privacy issues

[a] French administrative regulator in charge of data privacy.

Some examples

- Fiscal data
- Health data (potentially 200 To of highly sensitive data)
- National statistics (Eurostat)
- Environment

$\Rightarrow$ 600 researchers located in Europe are working on these data.

The CASD technology

The users

Insee

eurostat
EUROPEAN COMMISSION

GENERALI

DIRECTION GÉNÉRALE DES FINANCES PUBLIQUES

LA BANQUE POSTALE

Rte
Réseau de transport d'électricité

bpifrance

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE
DE L'AGRICULTURE
DE L'AGROALIMENTAIRE
ET DE LA FORÊT

Inserm
Institut national
de la santé et de la recherche médicale

IRDES
Institut de recherche
et documentation en
économie de la santé

The CASD technology

Some projects

- London School of Economics, London (Distribution of high income households between France and UK)
- University of London, London (Labor market: working hours as an adjustment mechanism)
- CREST, Paris (Absenteeism of Math and French teachers and success at middle school certificate)
- Sciences Po, Paris (Impact of the ZUS, ZRU, ZFU systems)
- INSERM, Kremlin Bicetre (Determining social and economic factors for causes of deaths)
- University of Paris 1/Banque de France (Over-indebtedness and wages' evolution in response to the economic cycle)
- GREQAM, Marseille (Impact of the resort to the support granted to restaurant owners)
- HEC, Jouy-en-Josas (Factors of success in new business start-up)

The CASD technology

The CASD technology is now available for private companies (banks, asset managers, industrials):

- On-site use (data are located in the company)
- Hosting at GENES

`http://www.casd.eu/`

Machine Learning

- Machine Learning and Econometrics
- Some Lessons About Machine Learning

What can economists learn about big data?

"*I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?*" (Varian, 2009).

Big Data: New Tricks for Econometrics

"*I believe that these methods have a lot to offer and should be more widely known and used by economists. In fact, my standard advice to graduate students these days is go to the computer science department and take a class in machine learning*" (Varian, 2013).

$\Rightarrow$ **Machine learning vs econometrics (classical statistics)**

The new framework

Econometrics

The 3 Pillars:

- Economic theory
- Parametric model
- Statistical inference

The statistical tools:

- Linear regression, Maximum likelihood & GMM
- Logit, Probit, Tobit, etc.
- ARMA, VaR, Cointegration, VECM, ARCH

$\Rightarrow$ Parsimony, goodness of fit, theoretical consistency, etc.

Machine learning

The 3 Pillars:

- Data & features
- Non-parametric model
- Cross-validation

The statistical tools:

- Shrinkage regression (ridge, lasso, Lars, elastic net, spike, slab)
- Ensemble learning (boosting, bagging)
- Random forests, neural nets, support vector machines, deep learning, etc.

$\Rightarrow$ Training/test sets, sparsity, success rate, etc.

Econometrics = probability

- Linear regression:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + u$$

- The variables x_1, x_2 and x_3 are given by the theory (standard transformation = log, lag)
- Probability distribution: (Y, X) is a Gaussian vector
- R^2 is the appropriate statistic to assess the goodness of fit
- Most of studies are interested by $\hat{\beta}$ and not by $\hat{y}$!
- F test, t statistic, p value, Wald test, etc.

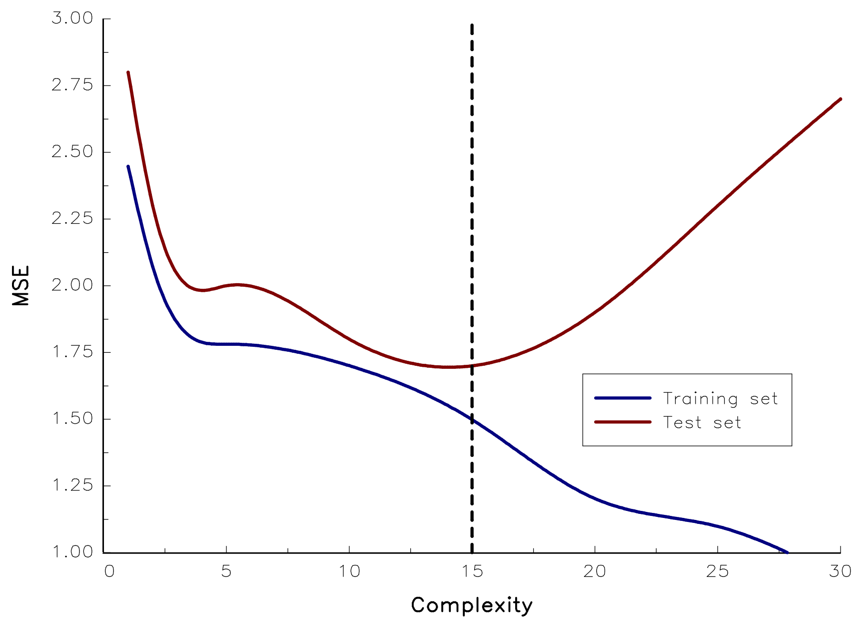
```
Valid cases:                  83     Dependent variable:           PUB6
Missing cases:                71     Deletion method:          Listwise
Total SS:               2027.807     Degrees of freedom:             80
R-squared:                 0.543     Rbar-squared:                0.531
Residual SS:             927.317     Std error of est:            3.405
F(2,80):                  47.470     Probability of F:            0.000

                          Standard                 Prob   Standardized  Cor with
Variable     Estimate       Error       t-value    >|t|     Estimate     Dep Var
-------------------------------------------------------------------------------
MMALE       -1.083609    1.455261    -0.744615    0.459       ---          ---
PUB3         0.800628    0.082676     9.683964    0.000    0.741746     0.709659
JOB          1.144342    0.437637     2.614821    0.011    0.200283     0.081449
```

Two lessons:

1. Focus on parameters!
2. The estimated model is **elegant**

Machine learning = a mix of statistics, computer science, economics, etc.

- Non-parametric model
- The important quantity is $\hat{y}$!
- It is not a probability model (inference statistics doesn't matter)
- Cross-validation step is very important (training set vs test set vs probe test)
- The solution may be **not elegant**



A famous big data competition

The Netflix competition (http://en.wikipedia.org/wiki/Netflix_Prize)

Netflix is a DVD rental/VOD company.

The Netflix prize (1 MUSD) was an open competition for the best collaborative filtering algorithm to predict user ratings for films, based on previous ratings.

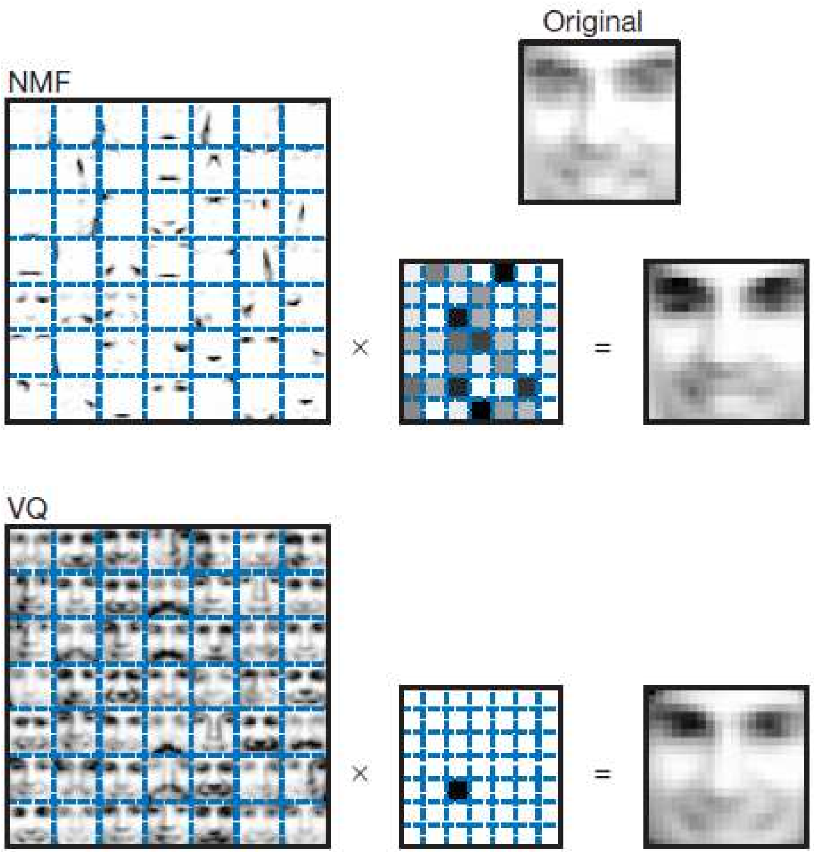
Training data = (user, movie, date, rating)
Test data = (user, movie, date)
Probe data = unknown

The winners are the joint team BellKor's Pragmatic Chaos
Solution = 154 pages to describe the model(s)!

A new way of thinking

Face recognition

- Human approach
  - Face: square, oval, circle
  - Hair: length, color, type
  - Typical patterns: eye, mouth, lips, ears, nose, eyelash, eyelid
  - Beard, mustache → male
  - Long hair → female
- Machine learning approaches
  - Holistic: vector quantization (VQ), k-NN
  - Semi-holistic: PCA, ICA, MAA, SVM
  - Parts-based approaches: NN, Non-negative matrix factorization (NMF)



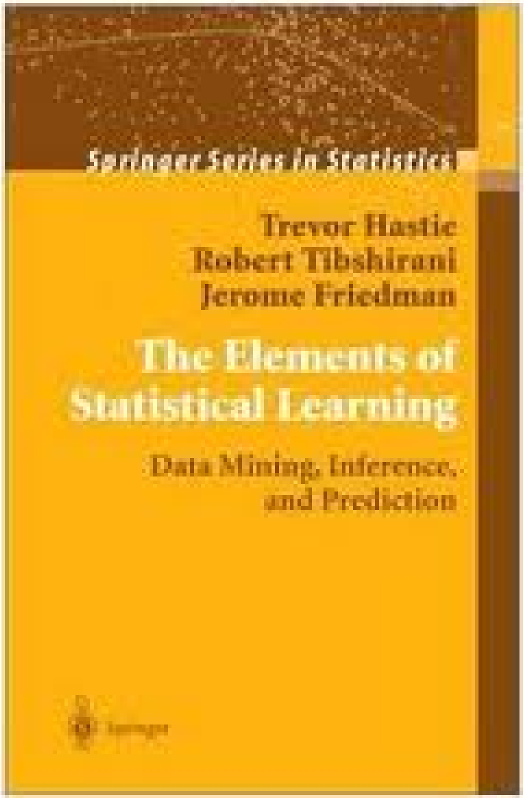
Source: Lee and Seung (1999).

Main discoveries

- Natural language processing (pattern recognition, neural networks)
- Classification (collaborative filtering, scoring, boosting, bagging, k-means)
- Lasso (loss function penalization)

Critical look about machine learning

- 2001: Publication of the Elements of Statistical Learning
- Frenzy impact on quants and HF managers ☺
- Disappointing results on investment strategies ☹

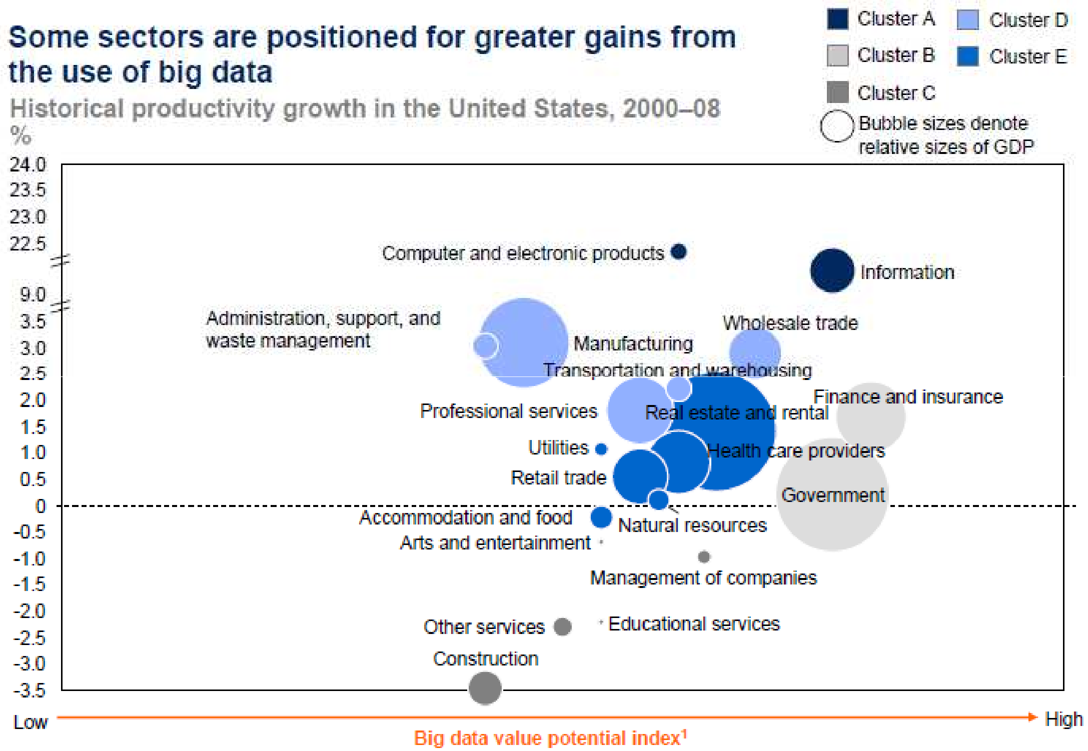


Some Applications in Asset Management

- Lasso Approach
- Nonnegative Matrix Factorization
- Measuring the Liquidity of ETFs
- The Art of Backtesting (or Data Mining)

High potential

According to McKinsey Global Institute, Finance & insurance "*are positioned to benefit very strongly from big data as long as barriers to its use can be overcome*".



Source: McKinsey Global Institute (2011).

Examples of big data problems in asset management

- Measuring the impact of high-frequency trading
- Measuring the liquidity
- Identification of systemic risks
- Computing the market portfolio
- Issues on collateral
- Etc.

Examples of big data problems in asset management

- and of course building trading strategies...

Blackrock Advisors UK Limited

Strategy: Pan European Equity market neutral strategy
Presented by: Richard Mathieson, Managing Director

Expertise:
Scientific Active Equities
- One of the pioneers of quantitative investing with almost two decades experience in running absolute return strategies
- A culture of continual innovation in combining investment insight with technology to deliver consistent and differentiated investment results for our clients.

Talking points:
Research driven investment process with growing application of signals constructed from big data:
- How we understand economic connections: macro research
- How new techniques enable us to capture more relevant information : text analysis
- How we identify less obvious linkages between stocks: clustering
- How we see the change in markets' structure; investors' trades can also point to opportunities: passive flows

Source: Lyxor Investment Conference, Grand Palais, 4 November 2014.

Lasso regression

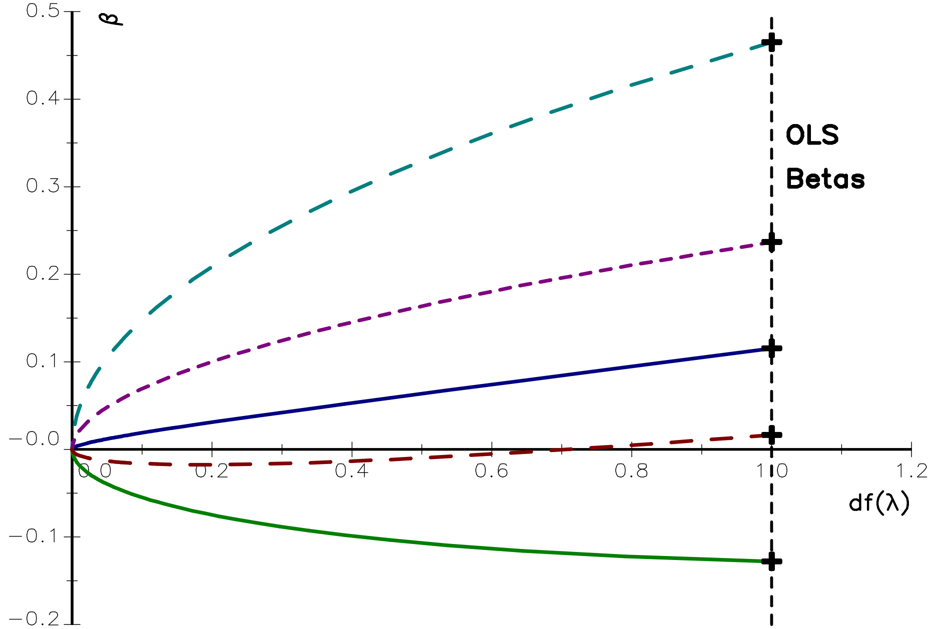
Lasso = a variant of the ridge regression with the L_1 norm penalty (Tibshirani, 1996):

$$\begin{aligned}\hat{\beta} &= \arg\min \left(Y - X\beta\right)^\top \left(Y - X\beta\right) \\ &\quad \text{u.c. } \sum_{j=1}^{m} \left|\beta_j\right| \leq \tau\end{aligned}$$

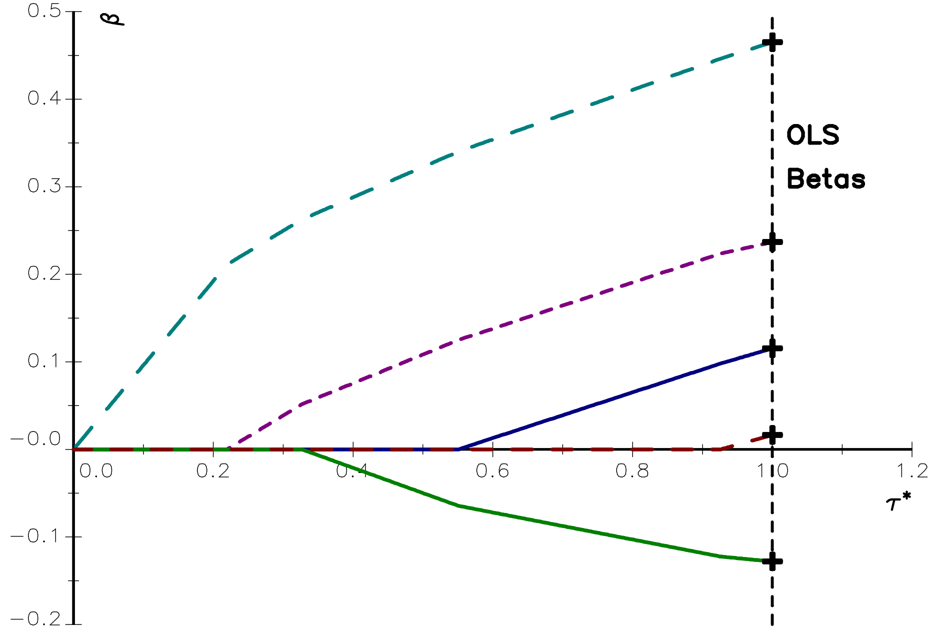
This problem is easy to solve using the quadratic programming framework and the parametrization $\beta = \beta^+ - \beta^-$ with $\beta^+ \geq \mathbf{0}$ and $\beta^- \geq \mathbf{0}$.

Lasso regression

Ridge / Tikhonov regularization



Lasso / L_1 regularization



Advantages of the lasso approach:

- Sparse model
- Selection model

Applications of the lasso approach

Here are some applications (Roncalli, 2014):

- Leverage
- Transaction costs
- Turnover
- Covariance matrix regularization
- Information matrix regularization
- Sparse portfolio optimization
- Sparse Kalman filtering

Hedge fund replication

1. Estimation step at time t

$$R_t^{\mathrm{HF}} = \sum_{i=1}^{m} \beta_{i,t} R_t^i + \varepsilon_t$$

where R_t^{HF} is the hedge funds return, R_t^i is the return of the i^{th} factor, $\beta_{i,t}$ is the exposure of hedge funds in the i^{th} factor and ε_t is a noise process.

2. Investment step for the time period $[t, t+1]$

$$R_{t+1}^{\mathrm{Tracker}} = \sum_{i=1}^{m} \hat{\beta}_{i,t} R_{t+1}^i$$

where $R_{t+1}^{\mathrm{Tracker}}$ is the return of the tracker.

The key issue is the selection of the factors (Roncalli and Weisang, 2009).

Hedge fund replication

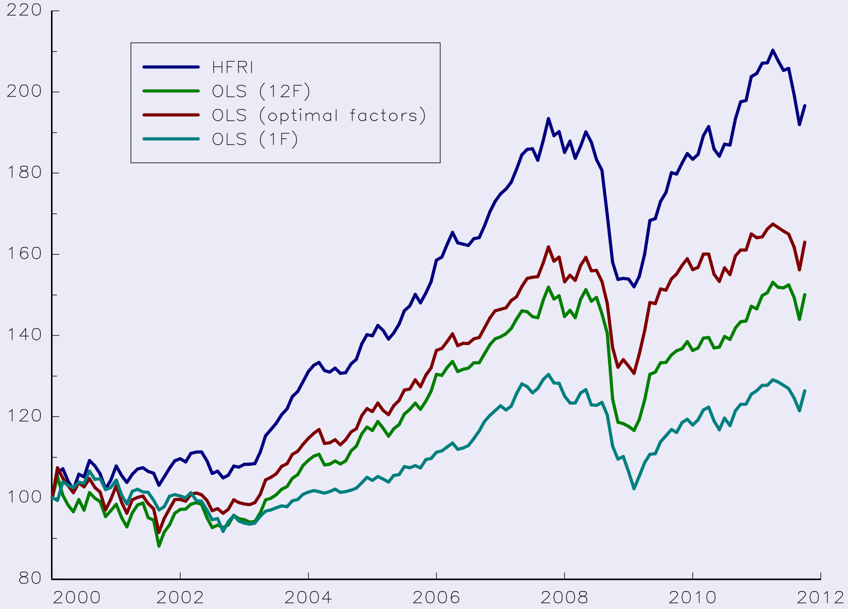
We consider the following set of 12 factors:

1. an equity exposure in the S&P 500 index (SPX)
2. a long/short position between Russell 2000 index and S&P5 00 index (RTY)
3. a long/short position between DJ Eurostoxx 50 index and S&P 500 index (SX5E)
4. a long/short position between TOPIX index and S&P 500 index (TPX)
5. a long/short position between MSCI EM index and S&P 500 index (MSCI EM)
6. an exposure in the 10Y US Treasury bond position (UST)
7. a FX position between Euro and US Dollar (EUR/USD)
8. a FX position between Yen and US Dollar (JPY/USD)
9. an exposure in high yield (HY)
10. an exposure in emerging bonds (EMBI)
11. an exposure in commodities (GSCI)
12. an exposure in gold (GOLD)

Hedge fund replication

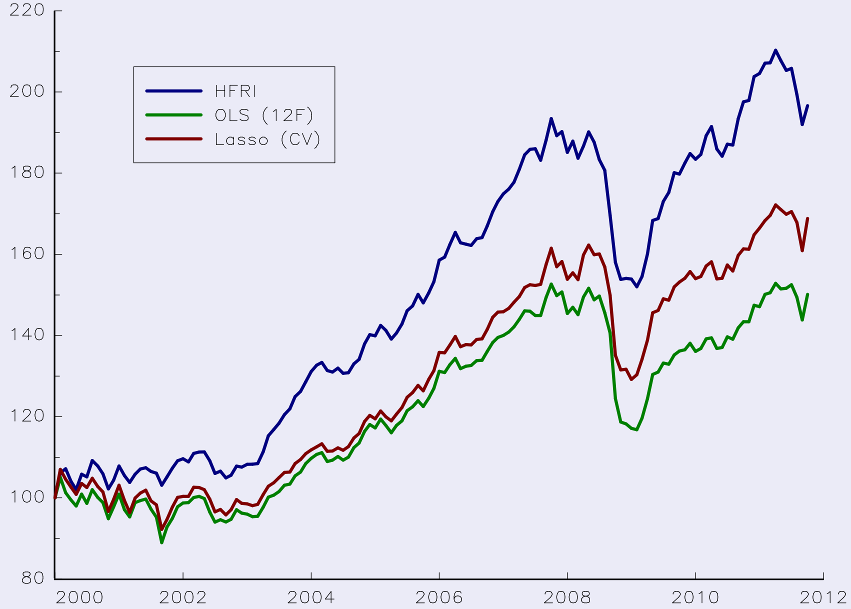
OLS tracker

- The performance depends on the definition of the universe of factors
- Choice of the factors = in-sample



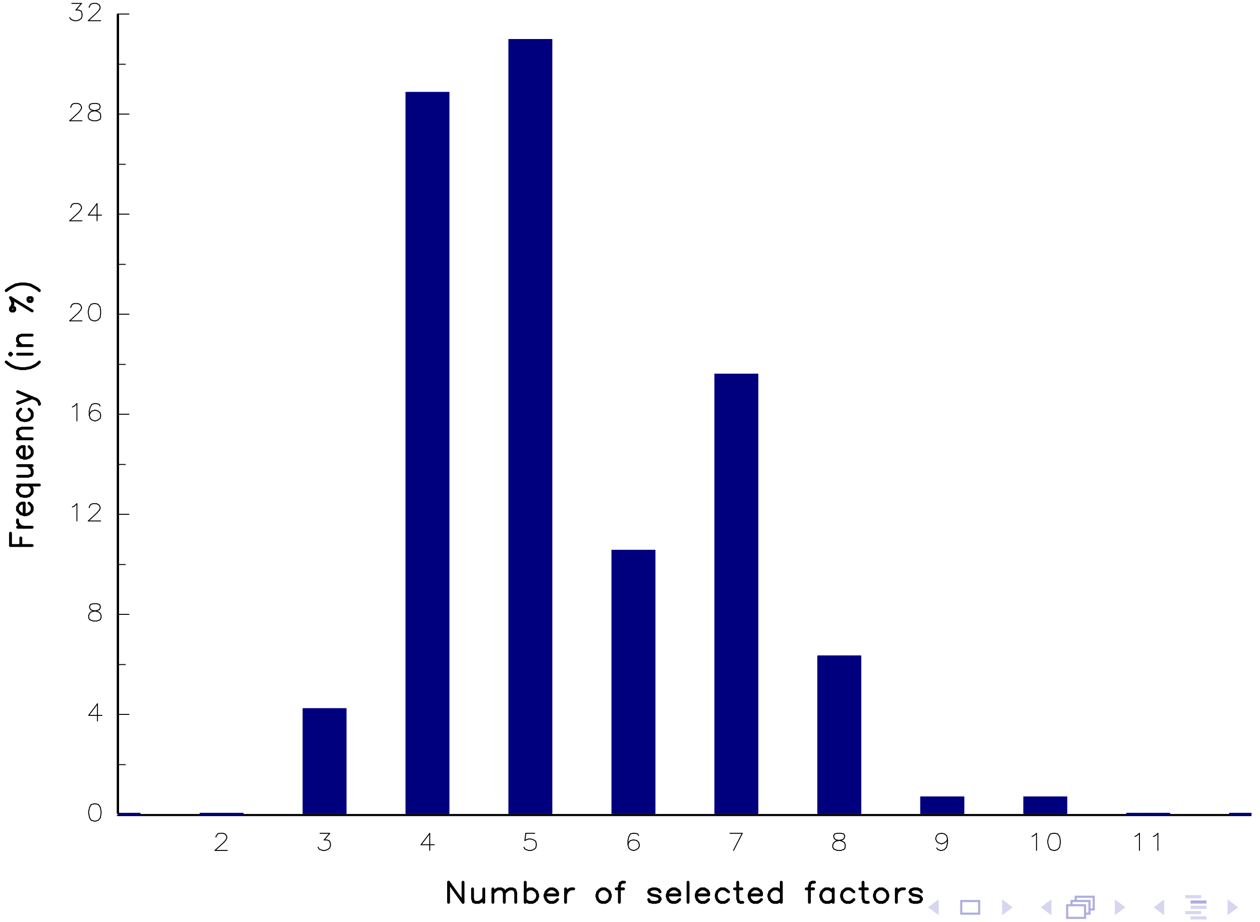
Lasso tracker

- Cross-validation based on the tracking error (Test set = 20%, 1000 bootstrap simulations)
- Choice of the factors = out-of-sample



Hedge fund replication

Figure: Number of selected factors (Lasso tracker, 2000-2012)



Portfolio optimization

The mean-variance optimization framework

We consider the following Markowitz optimization problem:

$$x^{\star}(\gamma) = \arg\min \frac{1}{2} x^{\top} \hat{\Sigma} x - \gamma x^{\top} \hat{\mu}$$

The solution is:

$$x^{\star}(\gamma) = \gamma \hat{\Sigma}^{-1} \hat{\mu}$$

The important quantity in mean-variance optimization is the information matrix $\mathcal{I} = \hat{\Sigma}^{-1}$.

Portfolio optimization

Interpretation of MVO portfolios

Stevens (1998) considers the following regression:

$$R_{i,t} = \beta_0 + \beta_i^\top R_t^{(-i)} + \varepsilon_{i,t}$$

where $R_t^{(-i)}$ denotes the vector of asset returns R_t excluding the i^{th} asset and $\varepsilon_{i,t} \sim \mathcal{N}(0, s_i^2)$. Stevens (1998) shows that:

$$\mathcal{I}_{i,i} = \frac{1}{\hat{\sigma}_i^2 \left(1 - R_i^2\right)}, \mathcal{I}_{i,j} = -\frac{\hat{\beta}_{i,j}}{\hat{\sigma}_i^2 \left(1 - R_i^2\right)} \text{ and } x_i^\star(\gamma) = \gamma \frac{\hat{\mu}_i - \hat{\beta}_i^\top \hat{\mu}^{(-i)}}{\hat{s}_i^2}$$

where $\hat{s}_i^2 = \hat{\sigma}_i^2 \left(1 - R_i^2\right)$. We deduce the following conclusions:

1. The better the hedge, the higher the exposure. This is why highly correlated assets produces unstable MVO portfolios.
2. The long-short position is defined by the sign of $\hat{\mu}_i - \hat{\beta}_i^\top \hat{\mu}^{(-i)}$. If the expected return of the asset is lower than the conditional expected return of the hedging portfolio, the weight is negative.

Portfolio optimization

Hedging portfolios with the empirical covariance matrix

Table: OLS hedging portfolios (in %) at the end of 2006

	SPX	SX5E	TPX	RTY	EM	US HY	EMBI	EUR	JPY	GSCI
SPX		58.6	6.0	150.3	-30.8	-0.5	5.0	-7.3	15.3	-25.5
SX5E	9.0		-1.2	-1.3	35.2	0.8	3.2	-4.5	-5.0	-1.5
TPX	0.4	-0.6		-2.4	38.1	1.1	-3.5	-4.9	-0.8	-0.3
RTY	48.6	-2.7	-10.4		26.2	-0.6	1.9	0.2	-6.4	5.6
EM	-4.1	30.9	69.2	10.9		0.9	4.6	9.1	3.9	33.1
US HY	-5.0	53.5	160.0	-18.8	69.5		95.6	48.4	31.4	-211.7
EMBI	10.8	44.2	-102.1	12.3	73.4	19.4		-5.8	40.5	86.2
EUR	-3.6	-14.7	-33.4	0.3	33.8	2.3	-1.4		56.7	48.2
JPY	6.8	-14.5	-4.8	-8.8	12.7	1.3	8.4	50.4		-33.2
GSCI	-1.1	-0.4	-0.2	0.8	10.7	-0.9	1.8	4.2	-3.3	
$\hat{s}_i$	0.3	0.7	0.9	0.5	0.7	0.1	0.2	0.4	0.4	1.2
R_i^2	83.0	47.7	34.9	82.4	60.9	39.8	51.6	42.3	43.7	12.1

Portfolio optimization

Hedging portfolios with the L_1 covariance matrix

Table: Lasso hedging portfolios (in %) at the end of 2006

	SPX	SX5E	TPX	RTY	EM	US HY	EMBI	EUR	JPY	GSCI
SPX		49.2		146.8			5.0	-3.2		
SX5E	5.1				32.3		3.2			
TPX					37.4	0.8	-3.1			
RTY	46.8		-3.1		10.4		1.9			
EM		25.0	61.3	6.5		0.8	4.3	2.6		22.4
US HY			82.2		65.9		93.8	19.1	20.7	
EMBI		24.9	-70.3		71.6	17.5			33.8	
EUR			-23.7		33.6	1.4			51.9	13.8
JPY					9.3	1.1	7.6	48.9		
GSCI					10.2	-0.3	1.6	2.9		
$\hat{s}_i$	0.3	0.7	1.0	0.5	0.7	0.1	0.2	0.4	0.4	1.3
R_i^2	82.0	44.9	33.9	82.1	60.4	38.0	51.5	39.7	41.5	7.8

Portfolio optimization

Backtest of the S&P 100 minimum-variance strategy

Bruder *et al.* (2013) consider the following backtest:

- Universe = S&P 100
- January 2000 - December 2011
- Monthly rebalancing

Table: Performance of OLS-MV and Lasso-MV portfolios

	$\mu(x)$	$\sigma(x)$	$\mathrm{SR}(x)$	$\mathcal{MDD}$	Turnover
OLS-MV	3.60%	14.39%	0.25	−39.71%	19.4
Lasso-MV	5.00%	13.82%	0.36	−35.42%	5.9

Remark

In December 2011, Google is hedged by 99 stocks if we consider the OLS-MV portfolio. Using the L_1 norm, Google is hedged by 13 stocks[a]*.*

[a] They are Boeing (4.6%), United technologies (1.1%), Schlumberger (1.8%), Williams cos. (1.8%), Microsoft (13.7%), Honeywell intl. (2.7%), Caterpillar (0.9%), Apple (25.0%), Mastercard (2.5%), Devon energy (2.9%), Nike (1.2%), Amazon (6.7%) and Apache (8.7%).

Nonnegative matrix factorization

Let A be a nonnegative matrix $m \times p$. We define the NMF decomposition as follows:

$$A = BC$$

where B and C are two nonnegative matrices with respective dimensions $m \times n$ and $n \times p$.

In the case where the objective function is to minimize the Frobenious norm, Lee and Seung (2001) propose to use the multiplicative update algorithm:

$$\begin{aligned} B_{(t+1)} &= B_{(t)} \odot \left(A C_{(t)}^\top \right) \oslash \left(B_{(t)} C_{(t)} C_{(t)}^\top \right) \\ C_{(t+1)} &= C_{(t)} \odot \left(B_{(t+1)}^\top A \right) \oslash \left(B_{(t+1)}^\top B_{(t+1)} C_{(t)} \right) \end{aligned}$$

where $\odot$ and $\oslash$ are respectively the element-wise multiplication and division operators. We have $\hat{B} = B_{(\infty)}$ and $\hat{C} = C_{(\infty)}$.

Interpretation of NMF

We consider the linear factor model $Y_t = \beta \mathcal{F}_t + \epsilon_t$ where Y_t is a $n \times 1$ vector and $\mathcal{F}_t$ is a $m \times 1$ vector. We note $Y = (Y_1, \ldots, Y_T)$ and $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_T)$.

- Principal component analysis

$$\Sigma = V \Lambda V^\top$$

  where Σ is the covariance matrix of $Y = (Y_1, \ldots, Y_T)$.

- Nonnegative matrix factorization:

$$A = BC$$

  where $A = Y^\top$, $B = \beta$ and $C = \mathcal{F}^\top$.

$\Rightarrow$ We may interpret B as a matrix of weights and C as a matrix of factors.

Remark

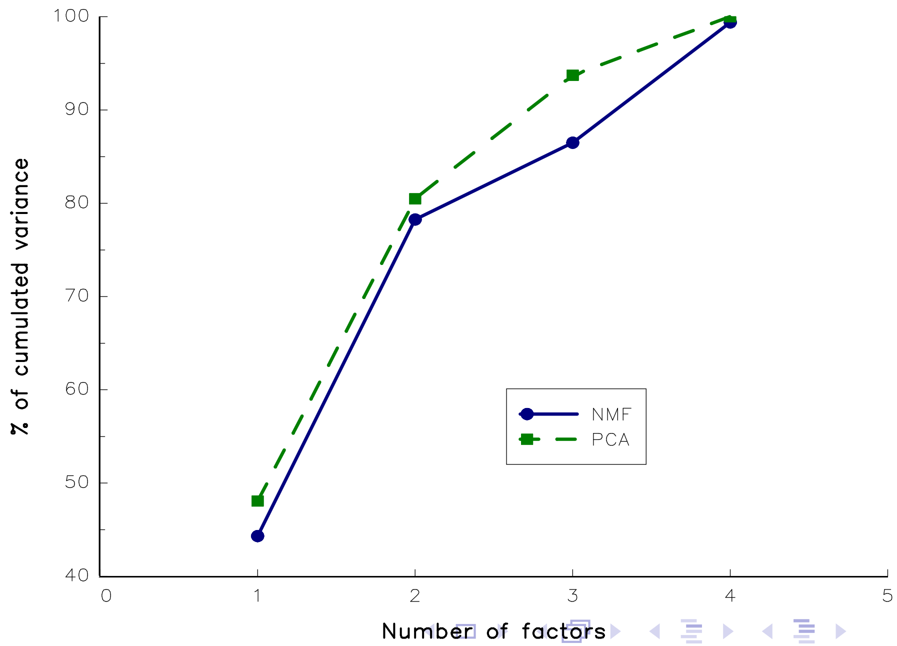
If $A = BC$, then $A^\top = C^\top B^\top$. The choice of the parametrization depends on the nature of the problem: analysis by observations (e.g. by trading dates) or by variables (e.g. by stocks).

Differences between NMF and PCA

- PCA: positive and negative weights = long-short portfolios
- NMF: positive weights = long-only portfolios ⇒ regularization & noise reduction

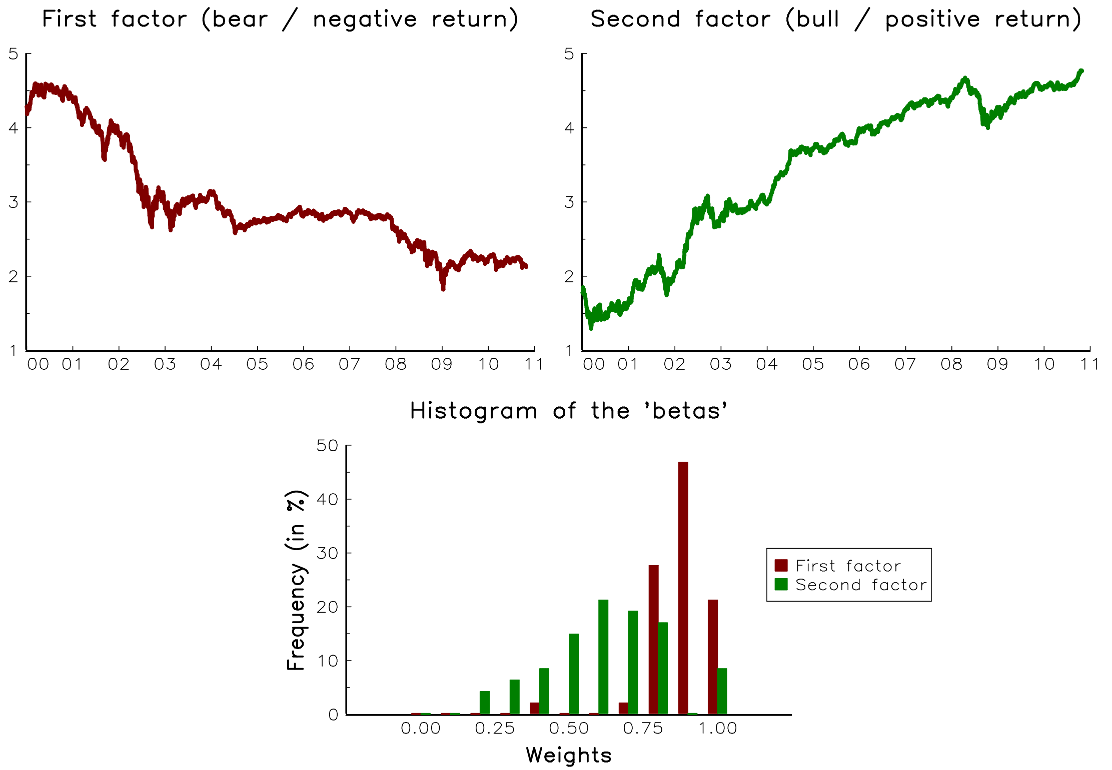
⇒ This implies that the cumulated variance explained by the first j PCA factors is always higher than the cumulated variance explained by the first j NMF factors.

An example with 4 stocks.



Sensitivity to bull/bear markets

Figure: NMF decomposition of the first 100 largest stocks of the EURO STOXX index



Stock classification

NMF may consider heterogeneous data (prices, volumes, Fama-French risk factors, etc.).

K-means classifier

	cluster										
sector	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	total
0001	5										5
1000	3				6			1			10
2000	4	2		1	7			1	1		16
3000	5	1			3	1			7		17
4000	4					1					5
5000	4			2	3			1			10
6000				6							6
7000	6										6
8000	4		8				3			4	19
9000		6									6
total	35	9	8	9	19	2	3	3	8	4	100

NMF classifier

	cluster										
sector	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	total
0001	1		2	0	1	1					5
1000	4	1	1		2	1			1		10
2000	3	1	3	1	1		2	4		1	16
3000	5			3	1	4	2		1	1	17
4000	1		3			1					5
5000		1	2	1		2			4		10
6000		3		3							6
7000			1		4				1		6
8000	3		2		1		3		4	6	19
9000		1		2					1	2	6
total	17	7	14	10	10	9	7	4	12	10	100

Sectors $\neq$ right clusters $\Rightarrow$ Some sectors are more heterogeneous than others (e.g. Financials: Banks, Insurance, Real Estate, Financial Services)

A big data problem?

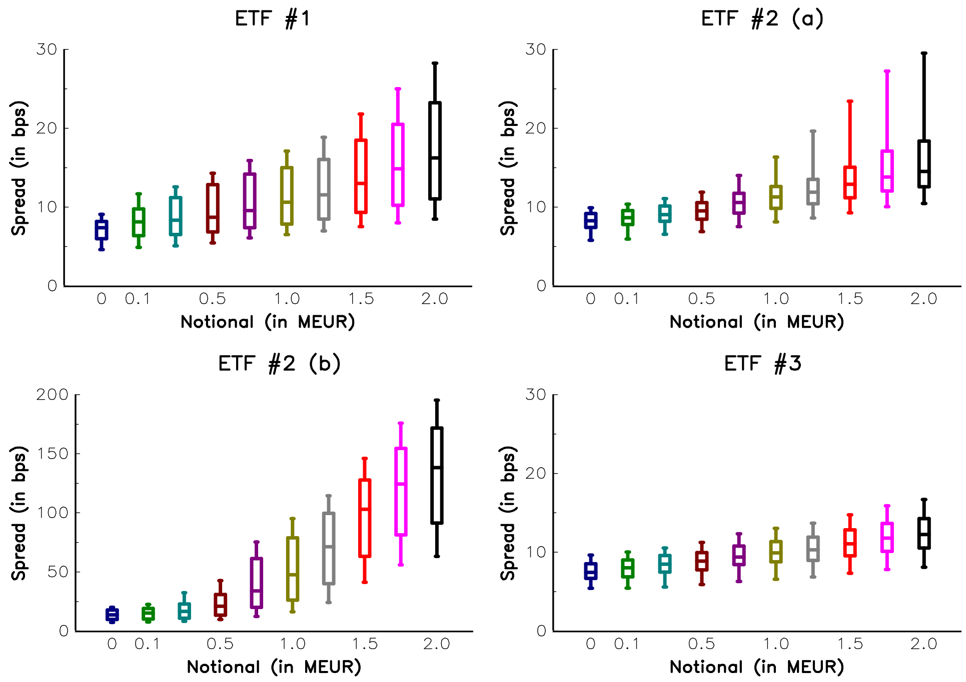
- Objective: measuring the liquidity of ETFs using limit order books in European markets.
- Difficulties:
  - 24 European exchanges
  - Cross-listing
  - Number of ETFs

Table: EURO STOXX 50 ETFs, 2012

ETF	Total card $(\mathcal{E})$	Total card $(\mathcal{T})$
#1	27 753 229	10 580
#2(a)	26 356 092	19 464
#2(b)	15 684 679	12 621
#3(a)	52 489 407	87 111
#3(b)	15 254 878	42 910
#4	50 423 195	95 539
Index	2 033 090 600	107 720 987

Computing the liquidity spread

Figure: Boxplot[2] of the liquidity spread (EURO STOXX 50, 2012)



[2]For each notional, the ends of the whiskers correspond to the minimum and maximum values. The bottom and top of the box are the 1st and 3rd quartiles whereas the median correspond to the line inside the box.

Computing the liquidity spread

Table: Median liquidity spread of MSCI World ETFs (in bps)

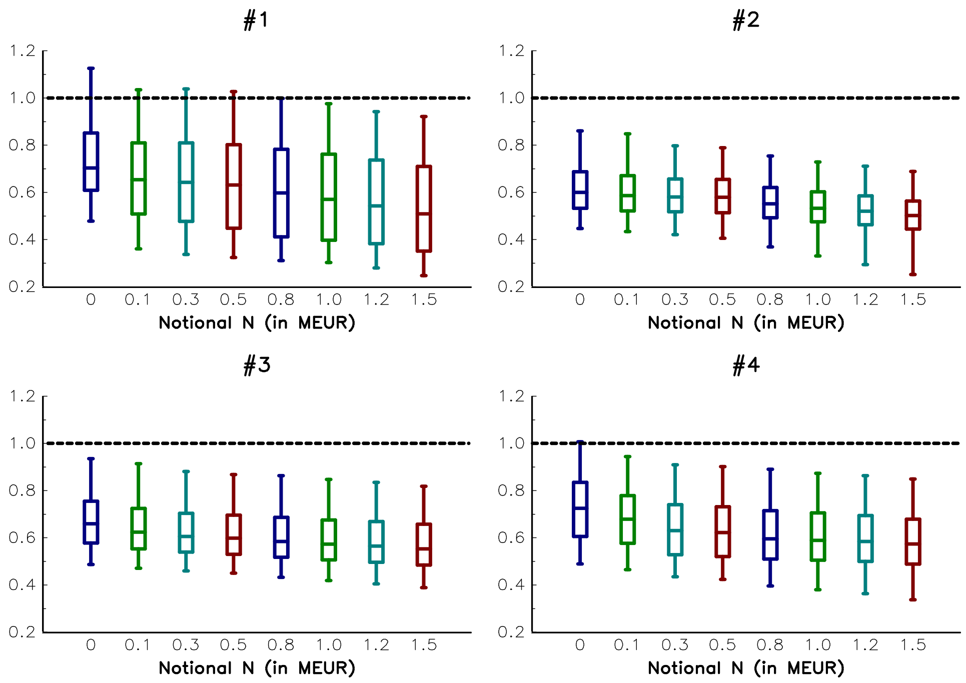
N (in MEUR)	#1	#2	#3	#4	#5	#6	#7	#8	#9
0.0	12	9	12	9	13	40	32	22	35
0.1	13	11	13	9	17	41	33	23	35
0.3	14	11	13	10	19	45	36	23	62
0.5	16	12	14	10	22	48	42	24	91
0.8	23	13	15	13	25	51	51	25	145
1.0	26	14	15	14	27	69	61	26	181
1.3	30	15	16	15	30	98	74	27	218
1.5	36	16	17	16	73	136	92	29	272
1.8	43	17	17	19	111	172	110	32	326
2.0	48	18	18	20	130	194	123	33	363

Liquidity spread may be highly different than the bid-ask spread.

Measuring the liquidity improvement

The liquidity ratio is defined as: $\mathcal{LR}_t(N) = \dfrac{S_t^{\mathrm{Index}}(N)}{S_t^{\mathrm{ETF}}(N)}$

Figure: Boxplot of the intraday liquidity ratio $\mathcal{LR}_t(N)$ (EURO STOXX 50)



What works / what doesn't

	Bond Scoring	Stock Picking	Trend Filtering	Mean Reverting	Index Tracking	HF Tracking	Stock Class.	Technical Analysis
Lasso		☺	☺	☺	☹	☺		
NMF							☺	☹
Boosting	☺	☺				☺		
Bagging	☺	☺				☺		
Random forests	☺			☹				☹
Neural nets	☺					☹		☹
SVM	☺	☹	☹				☹	☹
Sparse Kalman					☹	☺		
K-NN	☹							
K-means	☺						☺	
Testing protocols[3]	☺	☺	☺	☺		☺		

☺ = encouraging results

☹ = disappointing results

[3] Cross-validation, training/test/probe sets, K-fold, etc.

Backtesting and Sharpe ratio

- We consider a universe of n assets. Let μ and Σ be the vector of expected returns and the covariance matrix of asset returns. We note $x = (x_1, \ldots, x_n)$ the portfolio.
- The tangency portfolio is:

$$x^{\star} = \frac{\Sigma^{-1}(\mu - r\mathbf{1})}{\mathbf{1}^{\top}\Sigma^{-1}(\mu - r\mathbf{1})}$$

where r is the risk-free rate.

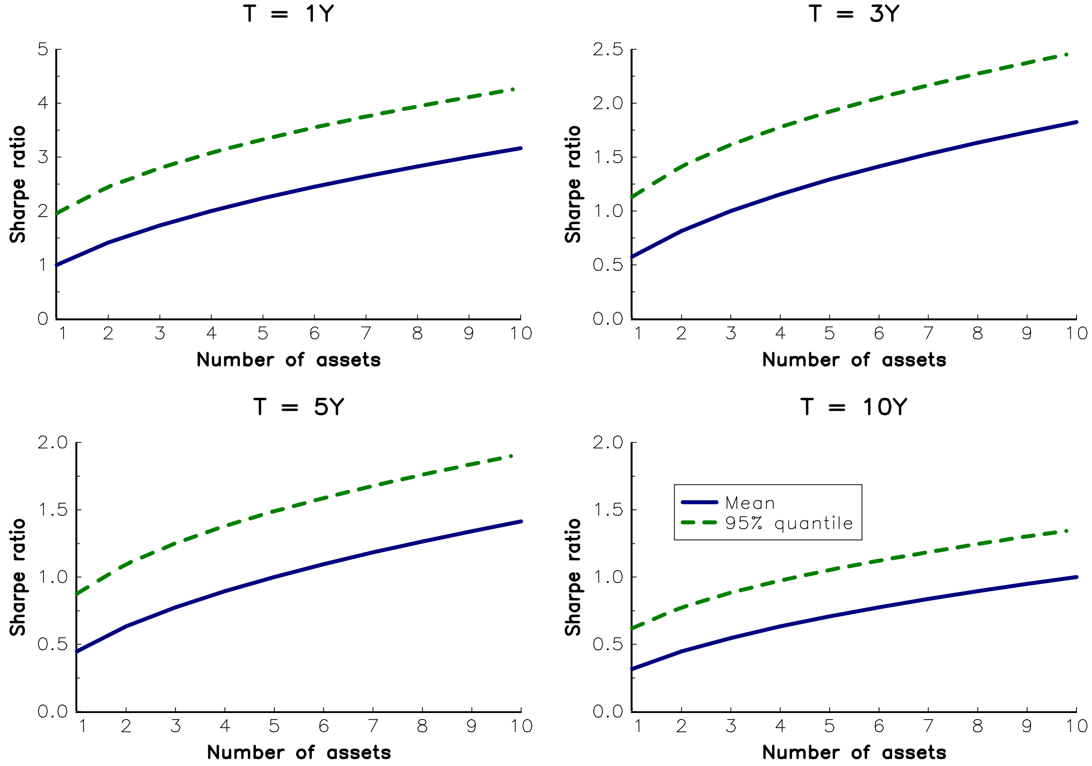
- If we consider the $\underline{\text{ex-post}}$ tangency portfolio of n correlated Brownian motions $(W_1(t), \ldots, W_n(t))$, we can show that:

$$\operatorname{sh}(x^{\star}; T) \sim \frac{\sqrt{\chi^2_n}}{\sqrt{T}}$$

where T is the time horizon used to estimate μ and Σ.

Backtesting and Sharpe ratio

Figure: Sharpe ratio of Markowitz portfolios



It is easy to find a blue line above the red line...

The example of the factor zoo

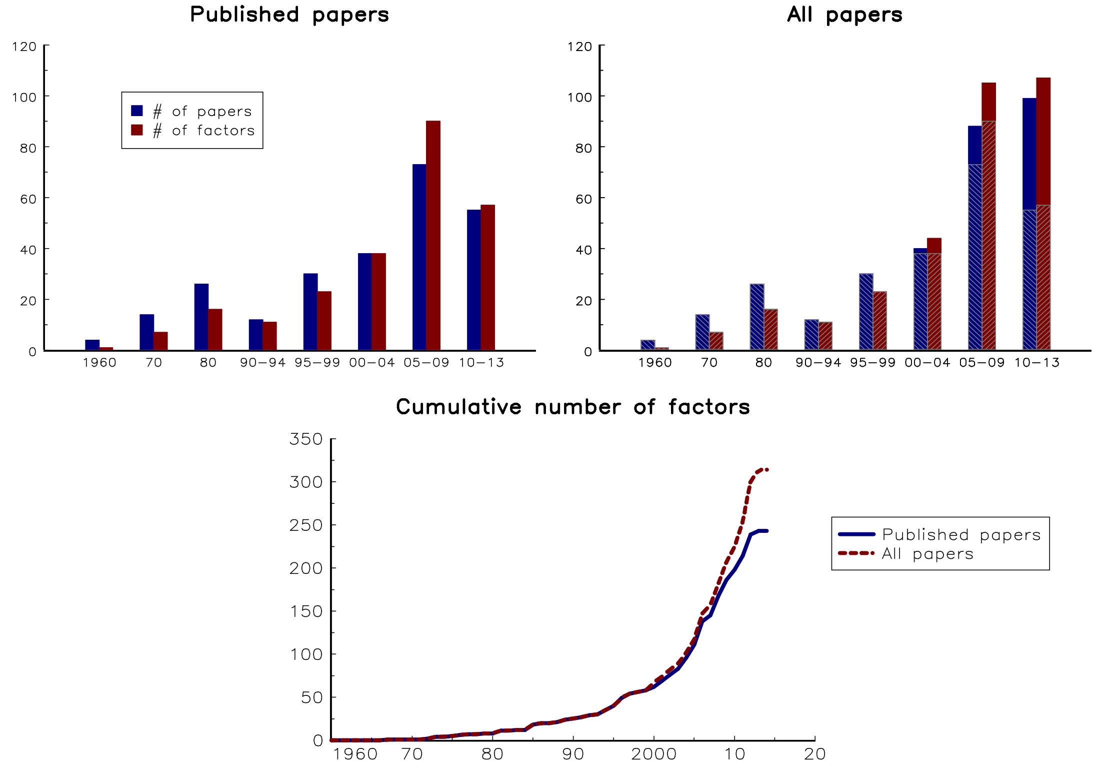
Factor investing

Factor investing is the second form of smart beta. It consists in investing in common risk factors (or new betas) that explain the variance of expected returns.

Examples of risk factors are: size, value, momentum, quality, short-term reversal, low beta, low volatility, liquidity, etc.

The example of the factor zoo

Figure: Harvey *et al.* (2014)



"*Now we have a zoo of new factors*" (Cochrane, 2011).

The example of the factor zoo

"Standard predictive regressions fail to reject the hypothesis that the party of the U.S. President, the weather in Manhattan, global warming, El Niño, sunspots, or the conjunctions of the planets, are significantly related to anomaly performance. These results are striking, and quite surprising. In fact, some readers may be inclined to reject some of this paper's conclusions solely on the grounds of plausibility. I urge readers to consider this option carefully, however, as doing do so entails rejecting the standard methodology on which the return predictability literature is built."(Novy-Marx, 2014).

$\Rightarrow$ **Do you think that they are risk factors or risk premia?**

False discoveries in biology and medicine

"90% of the world's data was created in the last two years" (IBM).

Conclusion

How Big Data can impact ESMA?

General I

The Economist.
Data, Data Everywhere.
February 2010.

McAfee A., Brynjolfsson E.
Big Data: The Management Revolution.
Harvard Business Review, 2012.

McKinsey Global Institute.
Big Data: The Next Frontier for Innovation, Competition and Productivity.
2011.

McKinsey Global Institute.
Disruptive Technologies: Advances that will Transform Life, Business, and the Global Economy.
2013.

General II

NATURE.
Big Data.
September 2008.

SCIENCE.
Dealing with Data.
February 2011.

VARIAN H.
Big Data: New Tricks for Econometrics.
SSRN, 2013.

The Netflix prize

Koren Y.
The BellKor Solution to the Netflix Grand Prize.
www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf, 2009.

Töscher A., Jahrer M., Bell R.
The BigChaos Solution to the Netflix Grand Prize.
www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf, 2009.

Piotte M., Chabbert M.
The Pragmatic Theory solution to the Netflix Grand Prize.
www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf, 2009.

Statistical learning I

BREIMAN L.
Bagging Predictors.
Machine Learning, 24, 1996.

EFRON B., HASTIE T., JOHNSTONE I., TIBSHIRANI R.
Least Angle Regression.
Annals of Statistics, 32(2), 2004.

FREUND Y.
Boosting a Weak Learning Algorithm by Majority.
Information and Computation, 121(2), 1995.

FREUND, Y. SHAPIRE R.E.
Experiments with a New Boosting Algorithm.
Machine Learning: Proceedings of the Thirteenth International Conference, Morgan Kaufman, 1996.

Statistical learning II

FRIEDMAN J.H.
Multivariate Adaptive Regression Splines.
Annals of Statistics, 19(1), 1991.

FRIEDMAN J.H., HASTIE T., TIBSHIRANI R.
Additive Logistic Regression: A Statistical View of Boosting.
Annals of Statistics, 28(2), 2000.

GUYON I., ELISSEEFF A.
An Introduction to Variable and Feature Selection.
Journal of Machine Learning Research, 3, 2003.

HASTIE T., TIBSHIRANI R., FRIEDMAN J.
The Elements of Statistical Learning.
Second edition, Springer, 2009.

LEE D.D., SEUNG H.S.
Learning the Parts of Objects by Non-negative Matrix Factorization.
Nature, 401, 1999.

Statistical learning III

SHAPIRE R.E.
The Strength of Weak Learnability.
Machine Learning, 5(2), 1990.

TIBSHIRANI R.
Regression Shrinkage and Selection via the Lasso.
Journal of the Royal Statistical Society B, 58(1), 1996.

TROPP J.A., GILBERT A.C.
Signal Recovery from Random Measurements via Orthogonal Matching Pursuit.
IEEE Transactions on Information Theory, 53(12), 2007.

VAPNIK V.
Statistical Learning Theory.
John Wiley and Sons, 1998.

ZOU H., HASTIE T., TIBSHIRANI R.
On the “Degrees of Freedom" of the Lasso.
Annals of Statistics, 35(5), 2007.

Backtesting I

COCHRANE J.H. (2011).
Presidential Address: Discount Rates.
Journal of Finance, 66(4), pp. 1047-1108.

HARVEY C.R., LIU Y. and ZHU H. (2014).
. . . and the Cross-Section of Expected Returns.
SSRN, www.ssrn.com/abstract=2249314

NOVY-MARX R. (2014).
Predicting Anomaly Performance with Politics, The Weather, Global Warming, Sunspots, and The Stars.
Journal of Financial Economics, 112(2), pp. 137-146.

CAZALET Z. and RONCALLI T. (2014).
Facts and Fantasies About Factor Investing.
Lyxor Research Paper, 112 pages.

Finance I

d'Aspremont A., Luss R.
Predicting Abnormal Returns From News Using Text Classification.
Quantitative Finance, 2012.

d'Aspremont A., Bach F., El Ghaoui L.
Optimal Solutions for Sparse Principal Component Analysis.
Journal of Machine Learning Research, 9, 2008.

Basak D., Pal S., Patranabis D.J.
Support Vector Regression.
Neural Information Processing, 11, 2007.

Belloni A., Chen D., Chernozhukov V., Hansen C.
Sparse Models and Methods for Optimal Instruments with An Application to Eminent Domain.
Econometrica, 80(6), 2012.

Finance II

Brodie J., Daubechies I., De Mol C., Giannone D., Loris I.
Sparse and Stable Markowitz Portfolios.
Proceedings of the National Academy of Sciences, 106(30), 2009.

Bruder B., Dao T-L., Richard J-C., Roncalli T.
Trend Filtering Methods for Momentum Strategies.
Lyxor White Paper Series, 8, 2011.

Bruder B., Gaussel N., Richard J-C., Roncalli T.
Regularization of Portfolio Allocation.
Lyxor White Paper Series, 10, 2013.

Carrasco M.
A Regularization Approach to the Many Instruments Problem.
Journal of Econometrics, 170(2), 2012.

Carrasco M., Noumon N.
Optimal Portfolio Selection using Regularization.
Working Paper, 2012.

Finance III

DeMiguel V., Garlappi L., Nogales F.J., Uppal R.
A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms.
Management Science, 55(5), 2009.

Fan J., Zhang J., Yu K.
Vast Portfolio Selection with Gross-exposure Constraints.
Journal of the American Statistical Association, 107(498), 2012.

Giamouridis D., Paterlini S.
Regular(ized) Hedge Fund Clones.
Journal of Financial Research, 33(3), 2010.

Kalaba R., Tesfatsion L.
Time-varying Linear Regression via Flexible Least Squares.
Computers & Mathematics with Applications, 17(8), 1989.

Finance IV

Kim S-J., Koh K., Boyd S., Gorinevsky D.
ℓ_1 Trend Filtering.
SIAM Review, 51(2), 2009.

Montana G., Triantafyllopoulos K., Tsagaris T.
Flexible Least Squares for Temporal Data Mining and Statistical Arbitrage.
Expert Systems with Applications, 36(2), 2009.

Roncalli T.
Introduction to Risk Parity and Budgeting.
Chapman & Hall, 2013.

Roncalli T., Weisang G.
Tracking Problems, Hedge Fund Replications and Alternative Beta.
Journal of Financial Transformation, 31, 2011.

Scherer B.
Portfolio Construction & Risk Budgeting.
Third edition, Risk Books, 2007.

Finance V

Sonneveld P., van Kan J.J.I.M., Huang X., Oosterlee C.W.
Nonnegative Matrix Factorization of a Correlation Matrix.
Linear Algebra and its Applications, 431(3-4), 2009.

Stevens G.V.G.
On the Inverse of the Covariance Matrix in Portfolio analysis.
Journal of Finance, 53(5), 1998.

Wu L., Yang Y., Liu H.
Nonnegative-Lasso and Application in Index Tracking.
Computational Statistics and Data Analysis, 70, 2013.

Yen Y.M., Yen T.J.
Solving Norm constrained Portfolio Optimization via Coordinate-wise Descent Algorithms.
Computational Statistics and Data Analysis, 2013.