Big Data in Asset Management

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

1The 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

1. What Does Big Data Mean?
   - Definition
   - Analysis of Big Data Problems
   - The Missing Factor

2. Data Management Issues
   - Some Examples
   - Managing Data Protection
   - Illustration with CASD

3. Machine Learning
   - Machine Learning and Econometrics

4. Some Lessons About Machine Learning

5. Some Applications in Asset Management
   - Lasso Approach
   - Nonnegative Matrix Factorization
   - Measuring the Liquidity of ETFs
   - The Art of Backtesting (or Data Mining)

5. Conclusion
What Does Big Data Mean?

- Definition
- Analysis of Big Data Problems
- The Missing Factor
What Does Big Data Mean?

A very hot topic

- Obama, White House, 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.

⇒ 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.)

⇒ Big data problems differ from one sector to another.

Big data ≠ 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.

An equal opportunity employer.
Figure: Supervised / unsupervised learning problems

- Large raw data
- Structured data
- Statistical learning
- Machine learning
- Fitting/Prediction model
- Classification model
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:
- + Veracity/Value

The challenge

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

1. Heterogeneous variables ⇒ comparable and workable data
2. Heterogeneous variables ⇒ new variables
3. Heterogeneous variables ⇒ valuable variables
4. Heterogeneous variables ⇒ model variables
The big challenge (1/2)

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!

Thierry Roncalli
Workshop on Big Data in Asset Management
11 / 80
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).

BIG DATA = DATA + ...
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).

⇒ Not simple.

How to complete this database with Newcits funds?

Who manages the project?

- IT?
- Experts?
The chicken-and-egg problem

An illustration

- 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)

⇒ 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
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

⇒ 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).

⇒ 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

Self-contained unit wholly dedicated to remote access

Smart Card Reader

Biometric Reader

Screen & On-site Setting Pushbuttons

RJ45 Connectors
VGA Connectors
USB Connectors
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 Configure
  - No Effects on the Rest of the IS
The CASD technology
The infrastructure

- A group of tightly-sealed secured servers
- User applications and processing are executed strictly within the Bubble
- Data Insertions/extractions are controlled. No Internet access from their workspace.
- Sensitive data is hosted only within the Bubble
- SD-Boxes are the only means of access to the Bubble.
- Access occurs via the Internet by encrypted channels
- Hadoop cluster is available for handling BigData

**Hermetic Bubble**

**Sensitive Data**

**Servers & Applications**
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

⇒ 600 researchers located in Europe are working on these data.
The CASD technology

The users
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 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).

⇒ 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

⇒ 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.

⇒ Training/test sets, sparsity, success rate, etc.
Econometrics \equiv 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.

| Variable | Estimate | Standard Error | t-value | Prob >|t| | Standardized Estimate | Cor with 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)!
Face recognition

- Human approach
  - Face: square, oval, circle
  - Hair: length, color, type
  - Typical patterns: eye, mouth, lips, ears, nose, eyelash, eyelid
  - Beard, mustache $\rightarrow$ male
  - Long hair $\rightarrow$ 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)

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)
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”.

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

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

$$\hat{\beta} = \arg\min (Y - X\beta)^T (Y - X\beta)$$

u.c. $\sum_{j=1}^{m} |\beta_j| \leq \tau$

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

Ridge / Tikhonov regularization

Lasso / $L_1$ regularization

Advantages of the lasso approach:
- Sparse model
- Selection model
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 $t$

$$ R_{t}^{HF} = \sum_{i=1}^{m} \beta_{i,t} R_{t}^{i} + \varepsilon_{t} $$

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 $\varepsilon_{t}$ is a noise process.

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

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

where $R_{t+1}^{Tracker}$ 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&P 500 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)
We consider the following Markowitz optimization problem:

\[ x^* (\gamma) = \arg \min_x \frac{1}{2} x^T \hat{\Sigma} x - \gamma x^T \hat{\mu} \]

The solution is:

\[ x^* (\gamma) = \gamma \hat{\Sigma}^{-1} \hat{\mu} \]

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

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

where $\hat{s}_i^2 = \hat{\sigma}_i^2 (1 - R_i^2)$. 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

<table>
<thead>
<tr>
<th></th>
<th>SPX</th>
<th>SX5E</th>
<th>TPX</th>
<th>RTY</th>
<th>EM</th>
<th>US HY</th>
<th>EMBI</th>
<th>EUR</th>
<th>JPY</th>
<th>GSCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPX</td>
<td>58.6</td>
<td>6.0</td>
<td>150.3</td>
<td>-30.8</td>
<td>-0.5</td>
<td>5.0</td>
<td>-7.3</td>
<td>15.3</td>
<td>-25.5</td>
<td></td>
</tr>
<tr>
<td>SX5E</td>
<td>9.0</td>
<td>-1.2</td>
<td>-1.3</td>
<td>35.2</td>
<td>0.8</td>
<td>3.2</td>
<td>-4.5</td>
<td>-5.0</td>
<td>-1.5</td>
<td></td>
</tr>
<tr>
<td>TPX</td>
<td>0.4</td>
<td>-0.6</td>
<td>-2.4</td>
<td>38.1</td>
<td>1.1</td>
<td>-3.5</td>
<td>-4.9</td>
<td>-0.8</td>
<td>-0.3</td>
<td></td>
</tr>
<tr>
<td>RTY</td>
<td>48.6</td>
<td>-2.7</td>
<td>-10.4</td>
<td>26.2</td>
<td>-0.6</td>
<td>1.9</td>
<td>0.2</td>
<td>-6.4</td>
<td>5.6</td>
<td></td>
</tr>
<tr>
<td>EM</td>
<td>-4.1</td>
<td>30.9</td>
<td>69.2</td>
<td>10.9</td>
<td>0.9</td>
<td>4.6</td>
<td>9.1</td>
<td>3.9</td>
<td>33.1</td>
<td></td>
</tr>
<tr>
<td>US HY</td>
<td>-5.0</td>
<td>53.5</td>
<td>160.0</td>
<td>-18.8</td>
<td>69.5</td>
<td>95.6</td>
<td>48.4</td>
<td>31.4</td>
<td>-211.7</td>
<td></td>
</tr>
<tr>
<td>EMBI</td>
<td>10.8</td>
<td>44.2</td>
<td>-102.1</td>
<td>12.3</td>
<td>73.4</td>
<td>19.4</td>
<td>-5.8</td>
<td>40.5</td>
<td>86.2</td>
<td></td>
</tr>
<tr>
<td>EUR</td>
<td>-3.6</td>
<td>-14.7</td>
<td>-33.4</td>
<td>0.3</td>
<td>33.8</td>
<td>2.3</td>
<td>-1.4</td>
<td>56.7</td>
<td>48.2</td>
<td></td>
</tr>
<tr>
<td>JPY</td>
<td>6.8</td>
<td>-14.5</td>
<td>-4.8</td>
<td>-8.8</td>
<td>12.7</td>
<td>1.3</td>
<td>8.4</td>
<td>50.4</td>
<td>-33.2</td>
<td></td>
</tr>
<tr>
<td>GSCI</td>
<td>-1.1</td>
<td>-0.4</td>
<td>-0.2</td>
<td>0.8</td>
<td>10.7</td>
<td>-0.9</td>
<td>1.8</td>
<td>4.2</td>
<td>-3.3</td>
<td></td>
</tr>
<tr>
<td>$\hat{s}_i$</td>
<td>0.3</td>
<td>0.7</td>
<td>0.9</td>
<td>0.5</td>
<td>0.7</td>
<td>0.1</td>
<td>0.4</td>
<td>0.4</td>
<td>1.2</td>
<td></td>
</tr>
<tr>
<td>$R^2_i$</td>
<td>83.0</td>
<td>47.7</td>
<td>34.9</td>
<td>82.4</td>
<td>60.9</td>
<td>39.8</td>
<td>51.6</td>
<td>42.3</td>
<td>43.7</td>
<td>12.1</td>
</tr>
</tbody>
</table>
### Table: Lasso hedging portfolios (in %) at the end of 2006

<table>
<thead>
<tr>
<th></th>
<th>SPX</th>
<th>SX5E</th>
<th>TPX</th>
<th>RTY</th>
<th>EM</th>
<th>US HY</th>
<th>EMBI</th>
<th>EUR</th>
<th>JPY</th>
<th>GSCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPX</td>
<td></td>
<td>49.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SX5E</td>
<td>5.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPX</td>
<td></td>
<td>146.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTY</td>
<td>46.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EM</td>
<td></td>
<td>25.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>US HY</td>
<td></td>
<td>61.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EMBI</td>
<td></td>
<td>37.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EUR</td>
<td></td>
<td>10.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>JPY</td>
<td></td>
<td>1.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GSCI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{s}_i$</td>
<td></td>
<td>0.3</td>
<td>0.7</td>
<td>1.0</td>
<td>0.5</td>
<td>0.7</td>
<td>0.1</td>
<td>0.2</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>$R^2_i$</td>
<td></td>
<td>82.0</td>
<td>44.9</td>
<td>33.9</td>
<td>82.1</td>
<td>60.4</td>
<td>38.0</td>
<td>51.5</td>
<td>39.7</td>
<td>41.5</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th></th>
<th>$\mu (x)$</th>
<th>$\sigma (x)$</th>
<th>SR$(x)$</th>
<th>MDD</th>
<th>Turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS-MV</td>
<td>3.60%</td>
<td>14.39%</td>
<td>0.25</td>
<td>−39.71%</td>
<td>19.4</td>
</tr>
<tr>
<td>Lasso-MV</td>
<td>5.00%</td>
<td>13.82%</td>
<td>0.36</td>
<td>−35.42%</td>
<td>5.9</td>
</tr>
</tbody>
</table>

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*.

\[ \text{a} \] 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%).
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:

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

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

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 $Y_t = \beta F_t + \epsilon_t$ where $Y_t$ is a $n \times 1$ vector and $F_t$ is a $m \times 1$ vector. We note $Y = (Y_1, \ldots, Y_T)$ and $F = (F_1, \ldots, F_T)$.

- Principal component analysis
  \[
  \Sigma = V \Lambda V^\top
  \]
  where $\Sigma$ is the covariance matrix of $Y = (Y_1, \ldots, Y_T)$.
- Nonnegative matrix factorization:
  \[
  A = BC
  \]
  where $A = Y^\top$, $B = \beta$ and $C = F^\top$.

$\Rightarrow$ 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.
**Figure:** NMF decomposition of the first 100 largest stocks of the EURO STOXX index

- **First factor (bear / negative return)**
- **Second factor (bull / positive return)**
NMF may consider heterogeneous data (prices, volumes, Fama-French risk factors, etc.).

<table>
<thead>
<tr>
<th>sector</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
<th>#7</th>
<th>#8</th>
<th>#9</th>
<th>#10</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>1000</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>2000</td>
<td>4</td>
<td>2</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>3000</td>
<td>5</td>
<td>1</td>
<td></td>
<td>3</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>4000</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>5000</td>
<td>4</td>
<td></td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>6000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>7000</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>8000</td>
<td>4</td>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td>19</td>
</tr>
<tr>
<td>9000</td>
<td></td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>total</td>
<td>35</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>19</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>8</td>
<td>4</td>
<td>100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sector</th>
<th>#1</th>
<th>#2</th>
<th>#3</th>
<th>#4</th>
<th>#5</th>
<th>#6</th>
<th>#7</th>
<th>#8</th>
<th>#9</th>
<th>#10</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>1</td>
<td></td>
<td>2</td>
<td>0</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>1000</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>2000</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
<td>2</td>
<td>4</td>
<td></td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>3000</td>
<td>5</td>
<td></td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>4000</td>
<td>1</td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>5000</td>
<td></td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td>2</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>6000</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
<tr>
<td>7000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>8000</td>
<td>3</td>
<td></td>
<td>2</td>
<td>1</td>
<td>3</td>
<td></td>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
<td>19</td>
</tr>
<tr>
<td>9000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>total</td>
<td>17</td>
<td>7</td>
<td>14</td>
<td>10</td>
<td>10</td>
<td>9</td>
<td>7</td>
<td>4</td>
<td>12</td>
<td>10</td>
<td>100</td>
</tr>
</tbody>
</table>

**Sectors ≠ right clusters** ⇒ Some sectors are more heterogeneous than others (e.g. Financials: Banks, Insurance, Real Estate, Financial Services)
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

<table>
<thead>
<tr>
<th>ETF</th>
<th>Total card (E)</th>
<th>Total card (T)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>27753229</td>
<td>10580</td>
</tr>
<tr>
<td>#2(a)</td>
<td>26356092</td>
<td>19464</td>
</tr>
<tr>
<td>#2(b)</td>
<td>15684679</td>
<td>12621</td>
</tr>
<tr>
<td>#3(a)</td>
<td>52489407</td>
<td>87111</td>
</tr>
<tr>
<td>#3(b)</td>
<td>15254878</td>
<td>42910</td>
</tr>
<tr>
<td>#4</td>
<td>50423195</td>
<td>95539</td>
</tr>
<tr>
<td>Index</td>
<td>2033090600</td>
<td>107720987</td>
</tr>
</tbody>
</table>
Figure: Boxplot\(^2\) of the liquidity spread (EURO STOXX 50, 2012)

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

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

Liquidity spread may be highly different than the bid-ask spread.
The liquidity ratio is defined as: $\mathcal{LR}_t(N) = \frac{S_t^{\text{Index}}(N)}{S_t^{\text{ETF}}(N)}$

**Figure:** Boxplot of the intraday liquidity ratio $\mathcal{LR}_t(N)$ (EURO STOXX 50)
### What works / what doesn’t

<table>
<thead>
<tr>
<th>Method</th>
<th>Bond Scoring</th>
<th>Stock Picking</th>
<th>Trend Filtering</th>
<th>Mean Reverting</th>
<th>Index Tracking</th>
<th>HF Tracking</th>
<th>Stock Class.</th>
<th>Technical Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NMF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boosting</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bagging</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random forests</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neural nets</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sparse Kalman</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-NN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K-means</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testing protocols</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 🎉 = 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 \( \mu \) and \( \Sigma \) 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:

\[
x^* = \frac{\Sigma^{-1} (\mu - r1)}{1^\top \Sigma^{-1} (\mu - r1)}
\]

where \( r \) is the risk-free rate.
- If we consider the ex-post tangency portfolio of \( n \) correlated Brownian motions \( (W_1(t), \ldots, W_n(t)) \), we can show that:

\[
sh(x^*; T) \sim \frac{\sqrt{\chi_n^2}}{\sqrt{T}}
\]

where \( T \) is the time horizon used to estimate \( \mu \) and \( \Sigma \).
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).
“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 so entails rejecting the standard methodology on which the return predictability literature is built.” (Novy-Marx, 2014).

⇒ Do you think that they are risk factors or risk premia?
“90% of the world’s data was created in the last two years” (IBM).
Conclusion

How Big Data can impact ESMA?
The Economist.
Data, Data Everywhere.
February 2010.

McAfee A., Brynjolfsson E.

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

McKinsey Global Institute.
2013.
General II

- **Nature.**
  Big Data.
  September 2008.

- **Science.**
  Dealing with Data.
  February 2011.

- **Varian H.**
  *SSRN*, 2013.
The Netflix prize

Koren Y.
The BellKor Solution to the Netflix Grand Prize.

Töscher A., Jahrer M., Bell R.
The BigChaos Solution to the Netflix Grand Prize.

Piotte M., Chabbert M.
The Pragmatic Theory solution to the Netflix Grand Prize.
Statistical learning I

Breiman L.
Bagging Predictors.

Efron B., Hastie T., Johnstone I., Tibshirani R.
Least Angle Regression.

Freund Y.
Boosting a Weak Learning Algorithm by Majority.

Freund, Y. Shapire R.E.
Experiments with a New Boosting Algorithm.
Statistical learning II

**Friedman J.H.**
Multivariate Adaptive Regression Splines.

**Friedman J.H., Hastie T., Tibshirani R.**
Additive Logistic Regression: A Statistical View of Boosting.

**Guyon I., Elisseeff A.**
An Introduction to Variable and Feature Selection.

**Hastie T., Tibshirani R., Friedman J.**
The Elements of Statistical Learning.

**Lee D.D., Seung H.S.**
Learning the Parts of Objects by Non-negative Matrix Factorization.
Statistical learning III

**Shapire R.E.**
The Strength of Weak Learnability.

**Tibshirani R.**
Regression Shrinkage and Selection via the Lasso.

**Tropp J.A., Gilbert A.C.**
Signal Recovery from Random Measurements via Orthogonal Matching Pursuit.

**Vapnik V.**
*Statistical Learning Theory*.

**Zou H., Hastie T., Tibshirani R.**
On the “Degrees of Freedom” of the Lasso.
Backtesting I

Presidential Address: Discount Rates.

...and the Cross-Section of Expected Returns.
*SSRN*, www.ssrn.com/abstract=2249314


Facts and Fantasies About Factor Investing.
d’Aspremont A., Luss R.
Predicting Abnormal Returns From News Using Text Classification.

d’Aspremont A., Bach F., El Ghaoui L.
Optimal Solutions for Sparse Principal Component Analysis.

Basak D., Pal S., Patranabis D.J.
Support Vector Regression.

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.
Some Applications in Asset Management

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.  

Bruder B., Gaussel N., Richard J-C., Roncalli T.  
Regularization of Portfolio Allocation.  

Carrasco M.  
A Regularization Approach to the Many Instruments Problem.  

Carrasco M., Noumon N.  
Optimal Portfolio Selection using Regularization.  
Finance III

DeMiguel V., Garlappi L., Nogales F.J., Uppal R.
A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms.

Fan J., Zhang J., Yu K.
Vast Portfolio Selection with Gross-exposure Constraints.

Giamouridis D., Paterlini S.
Regular(ized) Hedge Fund Clones.

Kalaba R., Tesfatsion L.
Time-varying Linear Regression via Flexible Least Squares.
What Does Big Data Mean?

Data Management Issues

Machine Learning

Some Applications in Asset Management

Conclusion

Finance IV

**Kim S-J., Koh K., Boyd S., Gorinevsky D.**

\( \ell_1 \) Trend Filtering.

*SIAM Review, 51(2), 2009.*

**Montana G., Triantafyllopoulos K., Tsagaris T.**

Flexible Least Squares for Temporal Data Mining and Statistical Arbitrage.


**Roncalli T.**

Introduction to Risk Parity and Budgeting.


**Roncalli T., Weisang G.**

Tracking Problems, Hedge Fund Replications and Alternative Beta.


**Scherer B.**

*Portfolio Construction & Risk Budgeting.*

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.

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.