Machine Learning Methods in Empirical Finance

Marcelo C. Medeiros

Departamento de Economia
Pontifícia Universidade Católica do Rio de Janeiro

Lecture 1
XVIII Encontro Brasileiro de Finanças

Introduction

- What is Machine Learning?

- What is Machine Learning?
  - What do we want to learn?

- What is Machine Learning?
  - What do we want to learn?
  - From what do we want to learn?

- What is Machine Learning?
  - What do we want to learn?
  - From what do we want to learn?
  - How do we want to learn?

What is Machine Learning (ML)?

- Automated computer **algorithms/methods** + **statistical models** to “learn” (discover) hidden patterns from data.

What is Machine Learning (ML)?

- Automated computer **algorithms/methods** + **statistical models** to "learn" (discover) hidden patterns from data.
- Usually ML methods are used for **prediction** (*prediction analytics*) but, more recently, they are also being applied to **causal inference**.

What is Machine Learning (ML)?

- Automated computer **algorithms/methods** + **statistical models** to “learn” (discover) hidden patterns from data.

- Usually ML methods are used for **prediction** (*prediction analytics*) but, more recently, they are also being applied to **causal inference**.

- ML methods are receiving a lot of attention in econometrics:

What is Machine Learning (ML)?

- Automated computer **algorithms/methods** + **statistical models** to “learn” (discover) hidden patterns from data.

- Usually ML methods are used for **prediction** (*prediction analytics*) but, more recently, they are also being applied to **causal inference**.

- ML methods are receiving a lot of attention in econometrics:

  – Model selection in data-rich environments (**big data**) for prediction and causal inference;

What is Machine Learning (ML)?

- Automated computer **algorithms/methods + statistical models** to "learn" (discover) hidden patterns from data.

- Usually ML methods are used for **prediction** (*prediction analytics*) but, more recently, they are also being applied to **causal inference**.

- ML methods are receiving a lot of attention in econometrics:
  - Model selection in data-rich environments (**big data**) for prediction and causal inference;
  - Nonlinear models.

What is Machine Learning (ML)?

- Automated computer **algorithms/methods** + **statistical models** to "learn" (discover) hidden patterns from data.

- Usually ML methods are used for **prediction** (*prediction analytics*) but, more recently, they are also being applied to **causal inference**.

- ML methods are receiving a lot of attention in econometrics:

  – Model selection in data-rich environments (**big data**) for prediction and causal inference;

  – Nonlinear models.

  – New inferential tools (post model selection).

What is Machine Learning (ML)?

- Automated computer **algorithms/methods +** **statistical models** to "learn" (discover) hidden patterns from data.

- Usually ML methods are used for **prediction** (*prediction analytics*) but, more recently, they are also being applied to **causal inference**.

- ML methods are receiving a lot of attention in econometrics:

  – Model selection in data-rich environments (**big data**) for prediction and causal inference;

  – Nonlinear models.

  – New inferential tools (post model selection).

- When ML methods are **statistically sound** they are called **Statistical Learning** (SL) methods.

What is Machine Learning (ML)?

ML versus Econometrics

Machine learning:

What is Machine Learning (ML)?

ML versus Econometrics

Machine learning:

- Main goal: prediction, classification, pattern recognition, cluster analysis, etc.

What is Machine Learning (ML)?

ML versus Econometrics

Machine learning:

- Main goal: prediction, classification, pattern recognition, cluster analysis, etc.
- Not much attention to inference or causal analysis, at least from a computer science perspective.

What is Machine Learning (ML)?

ML versus Econometrics

Machine learning:

- Main goal: prediction, classification, pattern recognition, cluster analysis, etc.
- Not much attention to inference or causal analysis, at least from a computer science perspective.
- Interpretation is not necessary a key ingredient.

What is Machine Learning (ML)?

ML versus Econometrics

Machine learning:

- Main goal: prediction, classification, pattern recognition, cluster analysis, etc.
- Not much attention to inference or causal analysis, at least from a computer science perspective.
- Interpretation is not necessary a key ingredient.
- Statistical learning gives more attention to inference and causal analysis.

What is Machine Learning (ML)?

ML versus Econometrics

Machine learning:

- Main goal: prediction, classification, pattern recognition, cluster analysis, etc.
- Not much attention to inference or causal analysis, at least from a computer science perspective.
- Interpretation is not necessary a key ingredient.
- Statistical learning gives more attention to inference and causal analysis.

Econometrics:

What is Machine Learning (ML)?

ML versus Econometrics

Machine learning:

- Main goal: prediction, classification, pattern recognition, cluster analysis, etc.
- Not much attention to inference or causal analysis, at least from a computer science perspective.
- Interpretation is not necessary a key ingredient.
- Statistical learning gives more attention to inference and causal analysis.

Econometrics:

- Statistical methods for prediction, inference, causal modeling of economic relationships.

What is Machine Learning (ML)?

ML versus Econometrics

Machine learning:

- Main goal: prediction, classification, pattern recognition, cluster analysis, etc.
- Not much attention to inference or causal analysis, at least from a computer science perspective.
- Interpretation is not necessary a key ingredient.
- Statistical learning gives more attention to inference and causal analysis.

Econometrics:

- Statistical methods for prediction, inference, causal modeling of economic relationships.
- Inference is a goal and interpretation is important.

What is Machine Learning (ML)?

ML versus Econometrics

Machine learning:

- Main goal: prediction, classification, pattern recognition, cluster analysis, etc.
- Not much attention to inference or causal analysis, at least from a computer science perspective.
- Interpretation is not necessary a key ingredient.
- Statistical learning gives more attention to inference and causal analysis.

Econometrics:

- Statistical methods for prediction, inference, causal modeling of economic relationships.
- Inference is a goal and interpretation is important.
- Causal inference is a goal for decision making.

A great matching:

Machine learning

with

Big Data

<u>A great matching:</u>

Machine learning

with

Big Data

with

Econometrics

What is "Big Data"?

"*The sexy job in the next ten years will be statisticians. Because now we really do have essentially free and ubiquitous data. So the complimentary factor is the ability to understand that data and extract value from it.*"

Hal Varian
Chief Economist, Google
January, 2009

What is “Big Data”?

“*The sexy job in the next ten years will be statisticians. Because now we really do have essentially free and ubiquitous data. So the complimentary factor is the ability to understand that data and extract value from it.*”

Hal Varian
Chief Economist, Google
January, 2009

- **Large** amount of data. We have data on everything!

What is "Big Data"?

"*The sexy job in the next ten years will be statisticians. Because now we really do have essentially free and ubiquitous data. So the complimentary factor is the ability to understand that data and extract value from it.*"

Hal Varian
Chief Economist, Google
January, 2009

- **Large** amount of data. We have data on everything!
- Large amount of variables and/or observations.

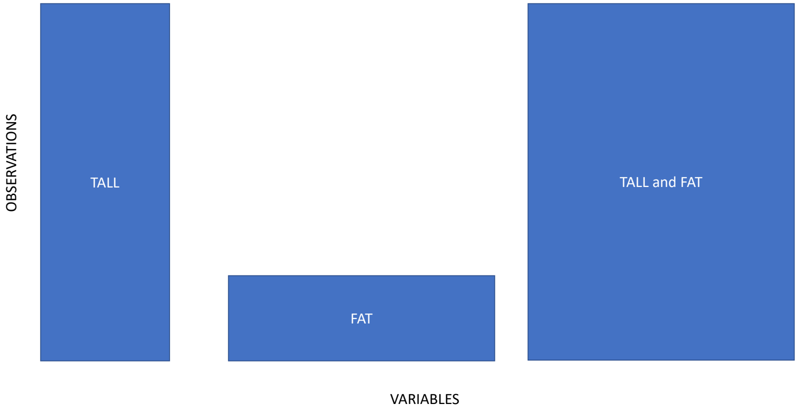
What is “Big Data”?

“*The sexy job in the next ten years will be statisticians. Because now we really do have essentially free and ubiquitous data. So the complimentary factor is the ability to understand that data and extract value from it.*”

Hal Varian
Chief Economist, Google
January, 2009

- **Large** amount of data. We have data on everything!
- Large amount of variables and/or observations.
- A quote from SAS (www.sas.comen_us/insightsbig-datawhat-is-big-data.html): “Big data is a term that describes the large volume of data – both **structured** and **unstructured** – that inundates a business on a day-to-day basis. But it’s not the amount of data that’s important. It’s what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.”

What is “Big Data”?



What is “Big Data”?

Structured versus unstructured data

Source: https://solutionsreview.com

What is “Big Data”?

Structured versus unstructured data

Structured data:

- **Highly organized** information that uploads nicely into traditional row database structures, lives in fixed fields, and is easily detectable via search operations or algorithms.

What is “Big Data”?

Structured versus unstructured data

Structured data:

- **Highly organized** information that uploads nicely into traditional row database structures, lives in fixed fields, and is easily detectable via search operations or algorithms.
- Is relatively simple to enter, store, query, and analyze, but it must be strictly defined in terms of field name and type (e.g. numeric, date, currency), and as a result is often restricted by character numbers or specific terminology.

What is “Big Data”?

Structured versus unstructured data

Structured data:

- **Highly organized** information that uploads nicely into traditional row database structures, lives in fixed fields, and is easily detectable via search operations or algorithms.
- Is relatively simple to enter, store, query, and analyze, but it must be strictly defined in terms of field name and type (e.g. numeric, date, currency), and as a result is often restricted by character numbers or specific terminology.

Unstructured data:

- **Everything else!**

What is “Big Data”?

Structured versus unstructured data

Structured data:

- **Highly organized** information that uploads nicely into traditional row database structures, lives in fixed fields, and is easily detectable via search operations or algorithms.
- Is relatively simple to enter, store, query, and analyze, but it must be strictly defined in terms of field name and type (e.g. numeric, date, currency), and as a result is often restricted by character numbers or specific terminology.

Unstructured data:

- **Everything else!**
- Unstructured data has internal structure but is not organized via pre-defined data models or schema.

What is “Big Data”?

Structured versus unstructured data

Structured data:

- **Highly organized** information that uploads nicely into traditional row database structures, lives in fixed fields, and is easily detectable via search operations or algorithms.
- Is relatively simple to enter, store, query, and analyze, but it must be strictly defined in terms of field name and type (e.g. numeric, date, currency), and as a result is often restricted by character numbers or specific terminology.

Unstructured data:

- **Everything else!**
- Unstructured data has internal structure but is not organized via pre-defined data models or schema.
- Examples: text files, web pages, social media, email, etc...

What is "Big Data"?

From unstructured to structured data

Example: Economic Policy Uncertainty

Baker, Bloom and Davies (QJE, 2016)

- Index from three types of underlying components:

What is “Big Data”?

From unstructured to structured data

Example: Economic Policy Uncertainty

Baker, Bloom and Davies (QJE, 2016)

- Index from three types of underlying components:
  1. First component quantifies **newspaper coverage** of policy-related economic uncertainty.

What is “Big Data”?

From unstructured to structured data

Example: Economic Policy Uncertainty

Baker, Bloom and Davies (QJE, 2016)

- Index from three types of underlying components:
    1. First component quantifies **newspaper coverage** of policy-related economic uncertainty.
    2. A second component reflects the number of federal tax code provisions set to expire in future years.

What is “Big Data”?

From unstructured to structured data

Example: Economic Policy Uncertainty

Baker, Bloom and Davies (QJE, 2016)

- Index from three types of underlying components:
  1. First component quantifies **newspaper coverage** of policy-related economic uncertainty.
  2. A second component reflects the number of federal tax code provisions set to expire in future years.
  3. The third component uses disagreement among economic forecasters as a proxy for uncertainty.

What is “Big Data”?

From unstructured to structured data

Example: Economic Policy Uncertainty

Baker, Bloom and Davies (QJE, 2016)

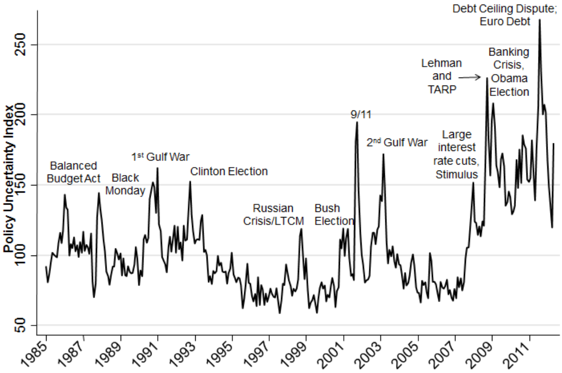
- Index from three types of underlying components:
  1. First component quantifies **newspaper coverage** of policy-related economic uncertainty.
  2. A second component reflects the number of federal tax code provisions set to expire in future years.
  3. The third component uses disagreement among economic forecasters as a proxy for uncertainty.
- From unstructured to structured data: The first component is an index of search results from 10 large newspapers. **Normalized index of the volume of news articles discussing economic policy uncertainty**.

  USA Today, the Miami Herald, the Chicago Tribune, the Washington Post, the Los Angeles Times, the Boston Globe, the San Francisco Chronicle, the Dallas Morning News, the New York Times, and the Wall Street Journal.

What is “Big Data”?

From unstructured to structured data

Example: Economic Policy Uncertainty



Source: http://www.policyuncertainty.com and Baker, Bloom and Davis(QJE, 2016).

What is “Big Data”?

From unstructured to structured data

Example: News implied VIX (NVIX)
Moreira and Manela (JFE, 2017)

- Text-based measure of uncertainty starting in 1890 using front-page articles of the Wall Street Journal.

What is “Big Data”?

From unstructured to structured data

Example: News implied VIX (NVIX)
Moreira and Manela (JFE, 2017)

- Text-based measure of uncertainty starting in 1890 using front-page articles of the Wall Street Journal.
- NVIX peaks during stock market crashes, times of policy-related uncertainty, world wars, and financial crises.

What is “Big Data”?

From unstructured to structured data

Example: News implied VIX (NVIX)
Moreira and Manela (JFE, 2017)

- Text-based measure of uncertainty starting in 1890 using front-page articles of the Wall Street Journal.
- NVIX peaks during stock market crashes, times of policy-related uncertainty, world wars, and financial crises.
- In US postwar data, periods when NVIX is high are followed by periods of above average stock returns, even after controlling for contemporaneous and forward-looking measures of stock market volatility.

What is "Big Data"?

From unstructured to structured data

Example: News implied VIX (NVIX)
Moreira and Manela (JFE, 2017)

- Text-based measure of uncertainty starting in 1890 using front-page articles of the Wall Street Journal.
- NVIX peaks during stock market crashes, times of policy-related uncertainty, world wars, and financial crises.
- In US postwar data, periods when NVIX is high are followed by periods of above average stock returns, even after controlling for contemporaneous and forward-looking measures of stock market volatility.
- NVIX is a key predictor of the equity premium.

What is “Big Data”?

From unstructured to structured data

Example: News implied VIX (NVIX)
Moreira and Manela (JFE, 2017)

- Text-based measure of uncertainty starting in 1890 using front-page articles of the Wall Street Journal.
- NVIX peaks during stock market crashes, times of policy-related uncertainty, world wars, and financial crises.
- In US postwar data, periods when NVIX is high are followed by periods of above average stock returns, even after controlling for contemporaneous and forward-looking measures of stock market volatility.
- NVIX is a key predictor of the equity premium.
- Methodology: ML regression of VIX on regressors based on text data.

What is “Big Data”?

From unstructured to structured data

Example: News implied VIX (NVIX)

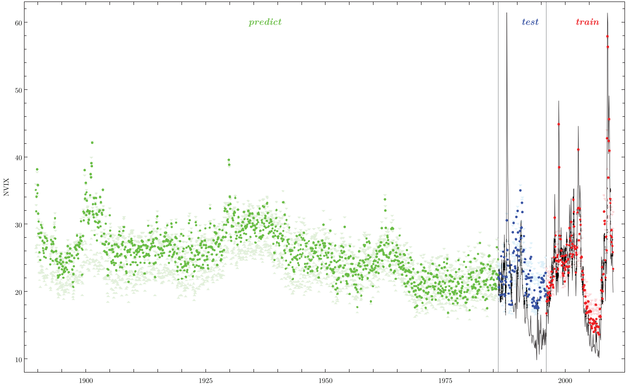
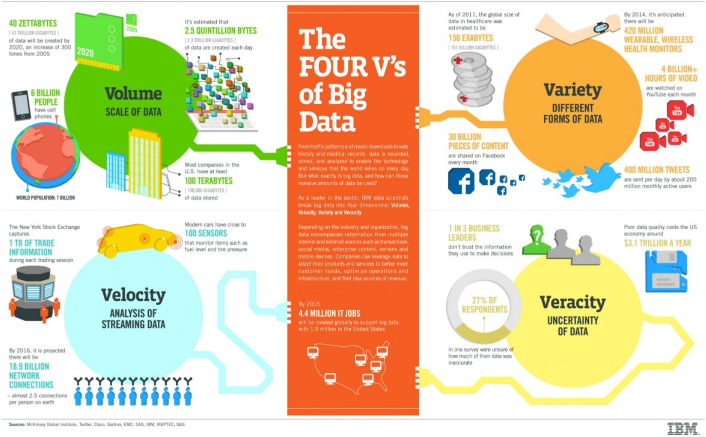


Fig. 1. News implied volatility 1890–2009. Solid line is end-of-month Chicago Board Options Exchange volatility implied by options VIX_t. Dots are news implied volatility (NVIX), $\widehat{VIX}_t = w_0 + \mathbf{w} \cdot \mathbf{x}_t$, where $x_{t,i}$ are appearances of n-gram i in month t scaled by total month t n-grams and $\mathbf{w}$ is estimated with a support vector regression. The *train* subsample, 1996 to 2009, is used to estimate the dependency between news data and implied volatility. The *test* subsample, 1986–1995, is used for out-of-sample tests of model fit. The *predict* subsample includes all earlier observations for which options data and, hence, VIX are not available. Light-colored triangles indicate a nonparametric bootstrap 95% confidence interval around $\widehat{VIX}$ using one thousand randomizations. These show the sensitivity of the predicted values to randomizations of the *train* subsample.

Source: Moreira and Manela (JFE, 2017).

What is “Big Data”?

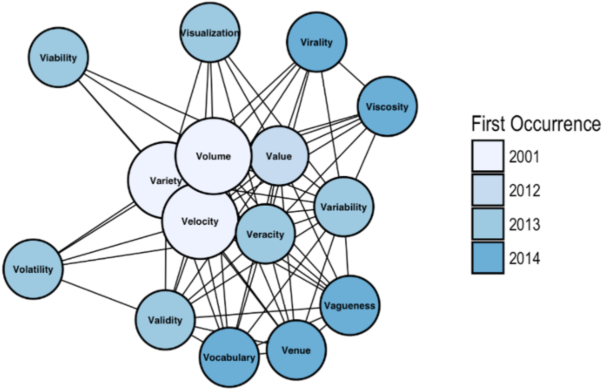
The Vs of “Big Data”



Source: http://www.ibmbigdatahub.com

What is “Big Data”?

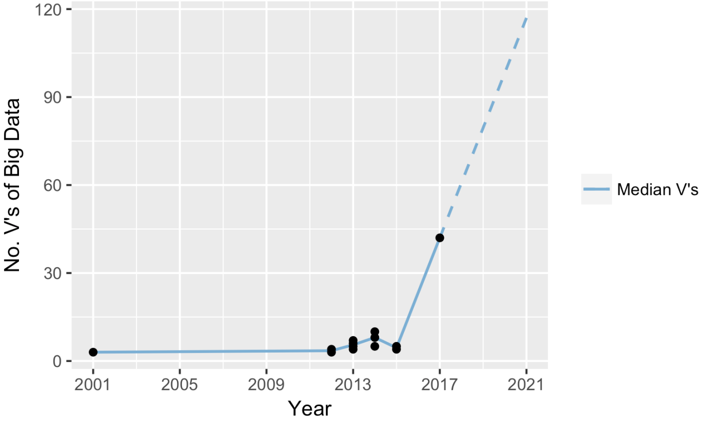
The Vs of “Big Data”



Source: https://www.elderresearch.com/company/blog/42-v-of-big-data

What is “Big Data”?

The Vs of “Big Data”



Source: https://www.elderresearch.com/company/blog/42-v-of-big-data

Machine Learning in Empirical Finance

- Equity premium forecasting
  Gu, Kelly and Xiu (2018)

Machine Learning in Empirical Finance

- Equity premium forecasting
  Gu, Kelly and Xiu (2018)
- Cross-section variability of stock returns/factor selection
  Bryzgalova (2015); Feng, Giglio and Xiu (2017); Kozak, Nagel and Santosh (2018), ...

Machine Learning in Empirical Finance

- Equity premium forecasting
  Gu, Kelly and Xiu (2018)
- Cross-section variability of stock returns/factor selection
  Bryzgalova (2015); Feng, Giglio and Xiu (2017); Kozak, Nagel and Santosh (2018), ...
- Covariance matrix forecast and portfolio choice
  Callot, Kock and Medeiros (JAE, 2017); Brito, Medeiros and Ribeiro (2018), ...

Machine Learning in Empirical Finance

- Equity premium forecasting
  Gu, Kelly and Xiu (2018)
- Cross-section variability of stock returns/factor selection
  Bryzgalova (2015); Feng, Giglio and Xiu (2017); Kozak, Nagel and Santosh (2018), ...
- Covariance matrix forecast and portfolio choice
  Callot, Kock and Medeiros (JAE, 2017); Brito, Medeiros and Ribeiro (2018), ...
- Volatility forecasting
  Scharth and Medeiros (IJF, 2009); Fernandes, Medeiros and Scharth (JBF, 2014), ...

Machine Learning in Empirical Finance

- Equity premium forecasting
  Gu, Kelly and Xiu (2018)
- Cross-section variability of stock returns/factor selection
  Bryzgalova (2015); Feng, Giglio and Xiu (2017); Kozak, Nagel and Santosh (2018), ...
- Covariance matrix forecast and portfolio choice
  Callot, Kock and Medeiros (JAE, 2017); Brito, Medeiros and Ribeiro (2018), ...
- Volatility forecasting
  Scharth and Medeiros (IJF, 2009); Fernandes, Medeiros and Scharth (JBF, 2014), ...
- Credit score, fraud detection, algorithmic trading, ...

Machine Learning in Empirical Finance

- Equity premium forecasting
  Gu, Kelly and Xiu (2018)
- Cross-section variability of stock returns/factor selection
  Bryzgalova (2015); Feng, Giglio and Xiu (2017); Kozak, Nagel and Santosh (2018), ...
- Covariance matrix forecast and portfolio choice
  Callot, Kock and Medeiros (JAE, 2017); Brito, Medeiros and Ribeiro (2018), ...
- Volatility forecasting
  Scharth and Medeiros (IJF, 2009); Fernandes, Medeiros and Scharth (JBF, 2014), ...
- Credit score, fraud detection, algorithmic trading, ...
- Lots of potential applications due to availability of massive datasets and new tools.

Models/Methods

What is a Machine Learning Model?

- One of the simplest ML method: **linear regression**!

$$\boxed{y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_p x_{pt} + u_t,\ t = 1, \ldots, T,}$$

where:

What is a Machine Learning Model?

- One of the simplest ML method: **linear regression**!

$$\boxed{y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_p x_{pt} + u_t,\ t = 1, \ldots, T,}$$

where:

  - y_t is the output variable (response) for element t, x_{jt}, $j = 1, \ldots, p$, is the j-th covariate for element t and u_t is the error term;

What is a Machine Learning Model?

- One of the simplest ML method: **linear regression**!

$$\boxed{y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_p x_{pt} + u_t,\ t = 1, \ldots, T,}$$

where:

– y_t is the output variable (response) for element t, x_{jt}, $j = 1, \ldots, p$, is the j-th covariate for element t and u_t is the error term;

– p parameters to be estimated $(\beta_1, \ldots, \beta_p)$ with T observations.

What is a Machine Learning Model?

- One of the simplest ML method: **linear regression**!

$$\boxed{y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_p x_{pt} + u_t,\ t = 1, \ldots, T,}$$

  where:
  - y_t is the output variable (response) for element t, x_{jt}, $j = 1, \ldots, p$, is the j-th covariate for element t and u_t is the error term;
  - p parameters to be estimated $(\beta_1, \ldots, \beta_p)$ with T observations.

- What do we learn?

What is a Machine Learning Model?

- One of the simplest ML method: **linear regression**!

$$\boxed{y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_p x_{pt} + u_t,\ t = 1, \ldots, T,}$$

  where:

  – y_t is the output variable (response) for element t, x_{jt}, $j = 1, \ldots, p$, is the j-th covariate for element t and u_t is the error term;

  – p parameters to be estimated $(\beta_1, \ldots, \beta_p)$ with T observations.

- What do we learn?

  – For sure: The best linear projection of y on the covariates $\boldsymbol{x} = (x_1, \ldots, x_p)'$. Exact solution by Ordinary Least Squares (OLS).

What is a Machine Learning Model?

- One of the simplest ML method: **linear regression**!

$$\boxed{y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_p x_{pt} + u_t,\ t = 1, \ldots, T,}$$

where:

– y_t is the output variable (response) for element t, x_{jt}, $j = 1, \ldots, p$, is the j-th covariate for element t and u_t is the error term;

– p parameters to be estimated $(\beta_1, \ldots, \beta_p)$ with T observations.

- What do we learn?

– For sure: The best linear projection of y on the covariates $\boldsymbol{x} = (x_1, \ldots, x_p)'$. Exact solution by Ordinary Least Squares (OLS).

– Under some assumptions: The $\mathbb{E}(y|\boldsymbol{x})$ or even the causal effects of changes in $\boldsymbol{x}$ on y

What is a Machine Learning Model?

- One of the simplest ML method: **linear regression**!

$$\boxed{y_t = \beta_0 + \beta_1 x_{1t} + \cdots + \beta_p x_{pt} + u_t,\ t = 1, \ldots, T,}$$

where:

  - y_t is the output variable (response) for element t, x_{jt}, $j = 1, \ldots, p$, is the j-th covariate for element t and u_t is the error term;
  - p parameters to be estimated $(\beta_1, \ldots, \beta_p)$ with T observations.

- What do we learn?
  - For sure: The best linear projection of y on the covariates $\boldsymbol{x} = (x_1, \ldots, x_p)'$. Exact solution by Ordinary Least Squares (OLS).
  - Under some assumptions: The $\mathbb{E}(y|\boldsymbol{x})$ or even the causal effects of changes in $\boldsymbol{x}$ on y

- Linear regression is a **GREAT** ML method!

What is a Machine Learning Model?

- However, in some cases, linear regression is not a good option:

What is a Machine Learning Model?

- However, in some cases, linear regression is not a good option:
  - High dimensions: $p > T \implies$ OLS is not feasible.

What is a Machine Learning Model?

- However, in some cases, linear regression is not a good option:
  - High dimensions: $p > T \implies$ OLS is not feasible.
  - Nonlinearities

What is a Machine Learning Model?

- However, in some cases, linear regression is not a good option:
  - High dimensions: $p > T \implies$ OLS is not feasible.
  - Nonlinearities
  - Nonlinearities + High dimensions

What is a Machine Learning Model?

- However, in some cases, linear regression is not a good option:
  - High dimensions: $p > T \implies$ OLS is not feasible.
  - Nonlinearities
  - Nonlinearities + High dimensions
- The cases above are becoming more and more frequent!

What is a Machine Learning Model?

- However, in some cases, linear regression is not a good option:
  - High dimensions: $p > T \implies$ OLS is not feasible.
  - Nonlinearities
  - Nonlinearities + High dimensions
- The cases above are becoming more and more frequent!
- Example: Moreira and Manela (JFE, 2017)

What is a Machine Learning Model?

- However, in some cases, linear regression is not a good option:
  - High dimensions: $p > T \Longrightarrow$ OLS is not feasible.
  - Nonlinearities
  - Nonlinearities + High dimensions
- The cases above are becoming more and more frequent!
- Example: Moreira and Manela (JFE, 2017)
  - y is the VIX and $\boldsymbol{x}$ is a vector with 468,091 entries representing one- and two-word n-gram frequencies from WJS frontpages:

$$x_{it} = \frac{\text{appearances of } n\text{-gram } i \text{ in month } t}{\text{total } n\text{-grams in month } t}$$

What is a Machine Learning Model?

- However, in some cases, linear regression is not a good option:
  - High dimensions: $p > T \implies$ OLS is not feasible.
  - Nonlinearities
  - Nonlinearities + High dimensions
- The cases above are becoming more and more frequent!
- Example: Moreira and Manela (JFE, 2017)
  - y is the VIX and $\boldsymbol{x}$ is a vector with 468,091 entries representing one- and two-word n-gram frequencies from WJS frontpages:

$$x_{it} = \frac{\text{appearances of } n\text{-gram } i \text{ in month } t}{\text{total } n\text{-grams in month } t}$$

  - An n-gram is a contiguous sequence of n items from a given sample of text or speech.

What is a Machine Learning Model?

- However, in some cases, linear regression is not a good option:
  - High dimensions: $p > T \Longrightarrow$ OLS is not feasible.
  - Nonlinearities
  - Nonlinearities + High dimensions
- The cases above are becoming more and more frequent!
- Example: Moreira and Manela (JFE, 2017)
  - y is the VIX and $\boldsymbol{x}$ is a vector with 468,091 entries representing one- and two-word n-gram frequencies from WJS frontpages:

$$x_{it} = \frac{\text{appearances of } n\text{-gram } i \text{ in month } t}{\text{total } n\text{-grams in month } t}$$

  - An n-gram is a contiguous sequence of n items from a given sample of text or speech.
  - The text is decomposed into five categories: War, Financial Intermediation, Government, Stock Markets, and Natural Disasters.

Machine Learning Methods

"*All models are wrong but some are useful.*"

George Box

- **New** ML models/methods/algorithms being proposed every day!

Machine Learning Methods

"*All models are wrong but some are useful.*"

George Box

- **New** ML models/methods/algorithms being proposed every day!
- **Old** models being rediscovered.

Machine Learning Methods

"*All models are wrong but some are useful.*"

George Box

- **New** ML models/methods/algorithms being proposed every day!
- **Old** models being rediscovered.
- Which model should we choose?

Machine Learning Methods

"*All models are wrong but some are useful.*"

George Box

- **New** ML models/methods/algorithms being proposed every day!
- **Old** models being rediscovered.
- Which model should we choose?
  - Linear versus nonlinear

Machine Learning Methods

"*All models are wrong but some are useful.*"

George Box

- **New** ML models/methods/algorithms being proposed every day!
- **Old** models being rediscovered.
- Which model should we choose?
  - Linear versus nonlinear
  - Parametric versus non-parametric versus semi-parametric

Machine Learning Methods

"*All models are wrong but some are useful.*"

George Box

- **New** ML models/methods/algorithms being proposed every day!
- **Old** models being rediscovered.
- Which model should we choose?
  - Linear versus nonlinear
  - Parametric versus non-parametric versus semi-parametric
  - Many different variable selection methods

Machine Learning Methods

"*All models are wrong but some are useful.*"

George Box

- **New** ML models/methods/algorithms being proposed every day!
- **Old** models being rediscovered.
- Which model should we choose?
  - Linear versus nonlinear
  - Parametric versus non-parametric versus semi-parametric
  - Many different variable selection methods
  - High risk of cherry picking!!! Data-mining in the bad sense of the term.

Machine Learning Methods

"*All models are wrong but some are useful.*"

George Box

- **New** ML models/methods/algorithms being proposed every day!
- **Old** models being rediscovered.
- Which model should we choose?
  - Linear versus nonlinear
  - Parametric versus non-parametric versus semi-parametric
  - Many different variable selection methods
  - High risk of cherry picking!!! Data-mining in the bad sense of the term.
- Names to keep in mind (just a few):

Machine Learning Methods

"*All models are wrong but some are useful.*"

George Box

- **New** ML models/methods/algorithms being proposed every day!
- **Old** models being rediscovered.
- Which model should we choose?
  - Linear versus nonlinear
  - Parametric versus non-parametric versus semi-parametric
  - Many different variable selection methods
  - High risk of cherry picking!!! Data-mining in the bad sense of the term.
- Names to keep in mind (just a few):
  - **Variable selection methods**: Bagging, Boosting, LASSO, Adaptive LASSO, Group LASSO, Fused LASSO, SCAD, Complete Subset Regression, Bayesian methods, factor models.

Machine Learning Methods

"*All models are wrong but some are useful.*"

George Box

- **New** ML models/methods/algorithms being proposed every day!
- **Old** models being rediscovered.
- Which model should we choose?
  - Linear versus nonlinear
  - Parametric versus non-parametric versus semi-parametric
  - Many different variable selection methods
  - High risk of cherry picking!!! Data-mining in the bad sense of the term.
- Names to keep in mind (just a few):
  - **Variable selection methods**: Bagging, Boosting, LASSO, Adaptive LASSO, Group LASSO, Fused LASSO, SCAD, Complete Subset Regression, Bayesian methods, factor models.
  - **Models**: linear regression, additive models, regression trees, random forests, neural networks, deep learning, kernel regression, series regression, splines.

Machine Learning Models

- **Supervised** versus **unsupervised** learning

Machine Learning Models

- **Supervised** versus **unsupervised** learning
- **Supervised learning**:

Machine Learning Models

- **Supervised** versus **unsupervised** learning
- **Supervised learning**:
  - input-output mapping:

$$\underbrace{y_t}_{\text{output}} = \underbrace{f}_{\text{mapping}} \underbrace{(\boldsymbol{x}_t)}_{\text{input (Big?)}} + \underbrace{u_t}_{\text{error}}$$

Machine Learning Models

- **Supervised** versus **unsupervised** learning
- **Supervised learning**:
  - input-output mapping:

$$\underbrace{y_t}_{\text{output}} = \underbrace{f}_{\text{mapping}} \underbrace{(\boldsymbol{x}_t)}_{\text{input (Big?)}} + \underbrace{u_t}_{\text{error}}$$

  - We need to choose the **vector of inputs** and the **mapping function**.

Machine Learning Models

- **Supervised** versus **unsupervised** learning
- **Supervised learning**:
  - input-output mapping:

$$\underbrace{y_t}_{\text{output}} = \underbrace{f}_{\text{mapping}} \underbrace{(\boldsymbol{x}_t)}_{\text{input (Big?)}} + \underbrace{u_t}_{\text{error}}$$

  - We need to choose the **vector of inputs** and the **mapping function**.
- **Unsupervised learning**:

Machine Learning Models

- **Supervised** versus **unsupervised** learning
- **Supervised learning**:
  - input-output mapping:

$$\underbrace{y_t}_{\text{output}} = \underbrace{f}_{\text{mapping}} \underbrace{(\boldsymbol{x}_t)}_{\text{input (Big?)}} + \underbrace{u_t}_{\text{error}}$$

  - We need to choose the **vector of inputs** and the **mapping function**.
- **Unsupervised learning**:
  - No inputs, just outputs!

Machine Learning Models

- **Supervised** versus **unsupervised** learning
- **Supervised learning**:
  - input-output mapping:

$$\underbrace{y_t}_{\text{output}} = \underbrace{f}_{\text{mapping}} \underbrace{(\boldsymbol{x}_t)}_{\text{input (Big?)}} + \underbrace{u_t}_{\text{error}}$$

  - We need to choose the **vector of inputs** and the **mapping function**.
- **Unsupervised learning**:
  - No inputs, just outputs!
  - The goal is to find "interesting" patters in data and there are no desired outputs given a set of inputs.

Machine Learning Models

- **Supervised** versus **unsupervised** learning
- **Supervised learning**:
  - input-output mapping:

$$\underbrace{y_t}_{\text{output}} = \underbrace{f}_{\text{mapping}} \underbrace{(\boldsymbol{x}_t)}_{\text{input (Big?)}} + \underbrace{u_t}_{\text{error}}$$

  - We need to choose the **vector of inputs** and the **mapping function**.
- **Unsupervised learning**:
  - No inputs, just outputs!
  - The goal is to find "interesting" patters in data and there are no desired outputs given a set of inputs.
  - Unconditional models, cluster analysis, missing value imputation, factor construction, etc.

Model Selection

- Back to the question: How should we choose a model?

Model Selection

- Back to the question: How should we choose a model?
  - Old forecasting school: choose the model with the best out-of-sample (OOS) performance.

Model Selection

- Back to the question: How should we choose a model?
  - Old forecasting school: choose the model with the best out-of-sample (OOS) performance.
  - Ensamble (forecast combination): use them all.

Model Selection

- Back to the question: How should we choose a model?
  - Old forecasting school: choose the model with the best out-of-sample (OOS) performance.
  - Ensamble (forecast combination): use them all.
  - Ensamble 2.0: use a subset of models.

Model Selection

- Back to the question: How should we choose a model?
  - Old forecasting school: choose the model with the best out-of-sample (OOS) performance.
  - Ensamble (forecast combination): use them all.
  - Ensamble 2.0: use a subset of models.
- This is still an open question!

Model Selection

- Back to the question: How should we choose a model?
  - Old forecasting school: choose the model with the best out-of-sample (OOS) performance.
  - Ensamble (forecast combination): use them all.
  - Ensamble 2.0: use a subset of models.
- This is still an open question!
- No free-lunch theorem (Wolpert, 1996): there is NO universal best model.

Model Selection

- Back to the question: How should we choose a model?
  - Old forecasting school: choose the model with the best out-of-sample (OOS) performance.
  - Ensamble (forecast combination): use them all.
  - Ensamble 2.0: use a subset of models.
- This is still an open question!
- No free-lunch theorem (Wolpert, 1996): there is NO universal best model.
  - The set of assumptions that works in one domain may work poorly in another.

Model Selection

- Back to the question: How should we choose a model?
  - Old forecasting school: choose the model with the best out-of-sample (OOS) performance.
  - Ensamble (forecast combination): use them all.
  - Ensamble 2.0: use a subset of models.
- This is still an open question!
- No free-lunch theorem (Wolpert, 1996): there is NO universal best model.
  - The set of assumptions that works in one domain may work poorly in another.
- Prediction versus causality.

Model Selection

- Back to the question: How should we choose a model?
  - Old forecasting school: choose the model with the best out-of-sample (OOS) performance.
  - Ensamble (forecast combination): use them all.
  - Ensamble 2.0: use a subset of models.
- This is still an open question!
- No free-lunch theorem (Wolpert, 1996): there is NO universal best model.
  - The set of assumptions that works in one domain may work poorly in another.
- Prediction versus causality.

$$\text{Big Data} + \text{Big Models} + \text{Big Set of Models}$$

$$=$$

BIG PROBLEM!!!!

Prediction and Inference after Model Selection

- Conducting inference with respect a set of parameters after model selection is a challenging task.

Prediction and Inference after Model Selection

- Conducting inference with respect a set of parameters after model selection is a challenging task.
- Finite sample inference is very complicated and the asymptotic results are usually not uniform over a wide class of probability distributions $\Rightarrow$ asymptotic distributions depend on the values of the true parameter.

Prediction and Inference after Model Selection

- Conducting inference with respect a set of parameters after model selection is a challenging task.
- Finite sample inference is very complicated and the asymptotic results are usually not uniform over a wide class of probability distributions $\Rightarrow$ asymptotic distributions depend on the values of the true parameter.
- Difficult to distinguish among smallish coefficients and zero.

Prediction and Inference after Model Selection

- Conducting inference with respect a set of parameters after model selection is a challenging task.
- Finite sample inference is very complicated and the asymptotic results are usually not uniform over a wide class of probability distributions $\Rightarrow$ asymptotic distributions depend on the values of the true parameter.
- Difficult to distinguish among smallish coefficients and zero.
- Inferential procedures must be adapted and conducting standard test ignoring model selection **is wrong**. Solution available for cross-section (see Victor Chernozhukov's papers). For time-series, solutions available only for specific settings; see Carvalho, Masini and Medeiros (JoE, in press).

Prediction and Inference after Model Selection

- Conducting inference with respect a set of parameters after model selection is a challenging task.
- Finite sample inference is very complicated and the asymptotic results are usually not uniform over a wide class of probability distributions $\Rightarrow$ asymptotic distributions depend on the values of the true parameter.
- Difficult to distinguish among smallish coefficients and zero.
- Inferential procedures must be adapted and conducting standard test ignoring model selection **is wrong**. Solution available for cross-section (see Victor Chernozhukov's papers). For time-series, solutions available only for specific settings; see Carvalho, Masini and Medeiros (JoE, in press).
- The lack of uniform convergence is not a problem of Big Data (high-dimensions) and it is due to the model search methods that are applied before inference is conducted.

Prediction and Inference after Model Selection

- Conducting inference with respect a set of parameters after model selection is a challenging task.
- Finite sample inference is very complicated and the asymptotic results are usually not uniform over a wide class of probability distributions $\Rightarrow$ asymptotic distributions depend on the values of the true parameter.
- Difficult to distinguish among smallish coefficients and zero.
- Inferential procedures must be adapted and conducting standard test ignoring model selection **is wrong**. Solution available for cross-section (see Victor Chernozhukov's papers). For time-series, solutions available only for specific settings; see Carvalho, Masini and Medeiros (JoE, in press).
- The lack of uniform convergence is not a problem of Big Data (high-dimensions) and it is due to the model search methods that are applied before inference is conducted.
- On the other hand, prediction (forecasting) after model selection is a much easier task.

Model Selection in High-Dimensions

- High-Dimensional Models:

Model Selection in High-Dimensions

- High-Dimensional Models:
  - **Relatively High-Dimension**: Models with many candidate variables p compared to the sample size n (or T), but usually less than n.

Model Selection in High-Dimensions

- High-Dimensional Models:
  - **Relatively High-Dimension**: Models with many candidate variables p compared to the sample size n (or T), but usually less than n.
  - **Moderately High-Dimension**: Models with candidate variables proportional to the sample size, usually greater than the sample size.

Model Selection in High-Dimensions

- High-Dimensional Models:
  - **Relatively High-Dimension**: Models with many candidate variables p compared to the sample size n (or T), but usually less than n.
  - **Moderately High-Dimension**: Models with candidate variables proportional to the sample size, usually greater than the sample size.
  - **High-Dimension**: Models with more candidate variables than observations, and the number of candidate variables grows polynomially or exponentially with n (or T).

Model Selection in High-Dimensions: Challenges

1. Prediction, oracle properties.
   Same prediction performance as the "true" model.

Model Selection in High-Dimensions: Challenges

1. Prediction, oracle properties.
   Same prediction performance as the "true" model.

2. Variable (Model) selection.
   Select <u>only</u> the correct set of relevant variables.

Model Selection in High-Dimensions: Challenges

1. Prediction, oracle properties.
   Same prediction performance as the “true” model.

2. Variable (Model) selection.
   Select <u>only</u> the correct set of relevant variables.

3. Variable screening.
   Select at least the correct set of variables.

Model Selection in High-Dimensions: Challenges

1. Prediction, oracle properties.
   Same prediction performance as the "true" model.

2. Variable (Model) selection.
   Select only the correct set of relevant variables.

3. Variable screening.
   Select at least the correct set of variables.

4. Inference.
   Distribution of the estimates.

Model Selection in High-Dimensions

- Estimation (model selection) in (linear) high dimension environments can be tackled in several ways:

Model Selection in High-Dimensions

- Estimation (model selection) in (linear) high dimension environments can be tackled in several ways:
  1. (Dynamic) Factor Models (DFM) $\Rightarrow$ dimension reduction.
     All variables are relevant but their variability can be summarized with a very small number of factors.

Model Selection in High-Dimensions

- Estimation (model selection) in (linear) high dimension environments can be tackled in several ways:
  1. (Dynamic) Factor Models (DFM) $\Rightarrow$ dimension reduction.
     All variables are relevant but their variability can be summarized with a very small number of factors.
  2. Penalized estimation (regularization)/shrinkage.
     Most of the variables are not relevant.

$$\text{cost} = \text{goodness of fit} + \text{penalty}.$$

Model Selection in High-Dimensions

- Estimation (model selection) in (linear) high dimension environments can be tackled in several ways:
  1. (Dynamic) Factor Models (DFM) $\Rightarrow$ dimension reduction.
     All variables are relevant but their variability can be summarized with a very small number of factors.
  2. Penalized estimation (regularization)/shrinkage.
     Most of the variables are not relevant.

     $$\text{cost} = \text{goodness of fit} + \text{penalty}.$$

  3. Bayesian methods (sort of shrinkage).

Model Selection in High-Dimensions

- Estimation (model selection) in (linear) high dimension environments can be tackled in several ways:
  1. (Dynamic) Factor Models (DFM) $\Rightarrow$ dimension reduction.
     All variables are relevant but their variability can be summarized with a very small number of factors.
  2. Penalized estimation (regularization)/shrinkage.
     Most of the variables are not relevant.

     $$\text{cost} = \text{goodness of fit} + \text{penalty}.$$

  3. Bayesian methods (sort of shrinkage).
  4. Bootstrap Aggregation (Bagging) and Boosting (sort of shrinkage).

Model Selection in High-Dimensions

- Estimation (model selection) in (linear) high dimension environments can be tackled in several ways:
  1. (Dynamic) Factor Models (DFM) $\Rightarrow$ dimension reduction.
     All variables are relevant but their variability can be summarized with a very small number of factors.
  2. Penalized estimation (regularization)/shrinkage.
     Most of the variables are not relevant.

     $$\textsf{cost} = \textsf{goodness of fit} + \textsf{penalty}.$$

  3. Bayesian methods (sort of shrinkage).
  4. Bootstrap Aggregation (Bagging) and Boosting (sort of shrinkage).
  5. Complete Subset Regressions (CSR)

Model Selection in High-Dimensions

- Estimation (model selection) in (linear) high dimension environments can be tackled in several ways:
  1. (Dynamic) Factor Models (DFM) $\Rightarrow$ dimension reduction.
     All variables are relevant but their variability can be summarized with a very small number of factors.
  2. Penalized estimation (regularization)/shrinkage.
     Most of the variables are not relevant.

$$\text{cost} = \text{goodness of fit} + \text{penalty}.$$

  3. Bayesian methods (sort of shrinkage).
  4. Bootstrap Aggregation (Bagging) and Boosting (sort of shrinkage).
  5. Complete Subset Regressions (CSR)
  6. Support Vector Machines (SVM)

Model Selection in High-Dimensions

- Estimation (model selection) in (linear) high dimension environments can be tackled in several ways:
  1. (Dynamic) Factor Models (DFM) $\Rightarrow$ dimension reduction.
     All variables are relevant but their variability can be summarized with a very small number of factors.
  2. Penalized estimation (regularization)/shrinkage.
     Most of the variables are not relevant.

$$\text{cost} = \text{goodness of fit} + \text{penalty}.$$

  3. Bayesian methods (sort of shrinkage).
  4. Bootstrap Aggregation (Bagging) and Boosting (sort of shrinkage).
  5. Complete Subset Regressions (CSR)
  6. Support Vector Machines (SVM)
- Nonlinear alternatives:

Model Selection in High-Dimensions

- Estimation (model selection) in (linear) high dimension environments can be tackled in several ways:
  1. (Dynamic) Factor Models (DFM) $\Rightarrow$ dimension reduction.
     All variables are relevant but their variability can be summarized with a very small number of factors.
  2. Penalized estimation (regularization)/shrinkage.
     Most of the variables are not relevant.

     $$\text{cost} = \text{goodness of fit} + \text{penalty}.$$

  3. Bayesian methods (sort of shrinkage).
  4. Bootstrap Aggregation (Bagging) and Boosting (sort of shrinkage).
  5. Complete Subset Regressions (CSR)
  6. Support Vector Machines (SVM)
- Nonlinear alternatives:
  1. Regression trees and neural networks

Model Selection in High-Dimensions

- Estimation (model selection) in (linear) high dimension environments can be tackled in several ways:
  1. (Dynamic) Factor Models (DFM) $\Rightarrow$ dimension reduction. All variables are relevant but their variability can be summarized with a very small number of factors.
  2. Penalized estimation (regularization)/shrinkage. Most of the variables are not relevant.

$$\text{cost} = \text{goodness of fit} + \text{penalty}.$$

  3. Bayesian methods (sort of shrinkage).
  4. Bootstrap Aggregation (Bagging) and Boosting (sort of shrinkage).
  5. Complete Subset Regressions (CSR)
  6. Support Vector Machines (SVM)
- Nonlinear alternatives:
  1. Regression trees and neural networks
  2. Shrinkage methods and bagging and boosting as well can be applied in nonlinear methods.

Model Selection in High-Dimensions

- Estimation (model selection) in (linear) high dimension environments can be tackled in several ways:
  1. (Dynamic) Factor Models (DFM) $\Rightarrow$ dimension reduction.
     All variables are relevant but their variability can be summarized with a very small number of factors.
  2. Penalized estimation (regularization)/shrinkage.
     Most of the variables are not relevant.

     $$\textsf{cost} = \textsf{goodness of fit} + \textsf{penalty}.$$

  3. Bayesian methods (sort of shrinkage).
  4. Bootstrap Aggregation (Bagging) and Boosting (sort of shrinkage).
  5. Complete Subset Regressions (CSR)
  6. Support Vector Machines (SVM)
- Nonlinear alternatives:
  1. Regression trees and neural networks
  2. Shrinkage methods and bagging and boosting as well can be applied in nonlinear methods.
  3. Bayesian methods.

The Road Map

Lecture 1:

- Linear models with shrinkage
- Applications to covariance matrix forecasting

Lecture 2:

- Nonlinear models
- Applications to equity premium forecasting

Shrinkage in Linear Models: Ridge, LASSO, Adaptive LASSO, Elastic Net

What happens when $p >> T$ in linear regressions?

Framework: Linear Regression Model

- We are interested in single-equation linear models

$$y_t = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_t + u_t$$

where

Framework: Linear Regression Model

- We are interested in single-equation linear models

$$y_t = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_t + u_t$$

where

  - $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})'$ is a vector of p exogenous variables,

Framework: Linear Regression Model

- We are interested in single-equation linear models

$$y_t = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_t + u_t$$

where

  – $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})'$ is a vector of p exogenous variables,
  – u_t is a zero-mean innovation,

Framework: Linear Regression Model

- We are interested in single-equation linear models

$$y_t = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_t + u_t$$

where

- $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})'$ is a vector of p exogenous variables,
- u_t is a zero-mean innovation,
- $\boldsymbol{x}_t = [\boldsymbol{x}_t(S)', \boldsymbol{x}_t(S^c)']'$, $\boldsymbol{x}_t(S) \in \mathbb{R}^s$ is the vector of **relevant** variables and $\boldsymbol{x}_t(S^c) \in \mathbb{R}^{p-s}$ is the vector of **irrelevant** ones. $\boldsymbol{\beta} = [\boldsymbol{\beta}_S', \boldsymbol{\beta}_{S^c}']'$.

Framework: Linear Regression Model

- We are interested in single-equation linear models

$$y_t = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_t + u_t$$

where

- $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})'$ is a vector of p exogenous variables,
- u_t is a zero-mean innovation,
- $\boldsymbol{x}_t = [\boldsymbol{x}_t(S)', \boldsymbol{x}_t(S^c)']'$, $\boldsymbol{x}_t(S) \in \mathbb{R}^s$ is the vector of **relevant** variables and $\boldsymbol{x}_t(S^c) \in \mathbb{R}^{p-s}$ is the vector of **irrelevant** ones. $\boldsymbol{\beta} = [\boldsymbol{\beta}'_S, \boldsymbol{\beta}'_{S^c}]'$.
- $p \equiv p(T)$ and $s \equiv s(T)$. T is the sample size.

Framework: Linear Regression Model

- We are interested in single-equation linear models

$$y_t = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_t + u_t$$

  where
  - $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})'$ is a vector of p exogenous variables,
  - u_t is a zero-mean innovation,
  - $\boldsymbol{x}_t = [\boldsymbol{x}_t(S)', \boldsymbol{x}_t(S^c)']'$, $\boldsymbol{x}_t(S) \in \mathbb{R}^s$ is the vector of **relevant** variables and $\boldsymbol{x}_t(S^c) \in \mathbb{R}^{p-s}$ is the vector of **irrelevant** ones. $\boldsymbol{\beta} = [\boldsymbol{\beta}_S', \boldsymbol{\beta}_{S^c}']'$.
  - $p \equiv p(T)$ and $s \equiv s(T)$. T is the sample size.
- Goals:

Framework: Linear Regression Model

- We are interested in single-equation linear models

$$y_t = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_t + u_t$$

  where

  – $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})'$ is a vector of p exogenous variables,
  – u_t is a zero-mean innovation,
  – $\boldsymbol{x}_t = [\boldsymbol{x}_t(S)', \boldsymbol{x}_t(S^c)']'$, $\boldsymbol{x}_t(S) \in \mathbb{R}^s$ is the vector of **relevant** variables and $\boldsymbol{x}_t(S^c) \in \mathbb{R}^{p-s}$ is the vector of **irrelevant** ones. $\boldsymbol{\beta} = [\boldsymbol{\beta}'_S, \boldsymbol{\beta}'_{S^c}]'$.
  – $p \equiv p(T)$ and $s \equiv s(T)$. T is the sample size.
- Goals:
  1. Select the right set of variables: $\widehat{\boldsymbol{\beta}}_S \neq \mathbf{0}$ and $\widehat{\boldsymbol{\beta}}_{S^c} = \mathbf{0}$ (model selection).

Framework: Linear Regression Model

- We are interested in single-equation linear models

$$y_t = \beta_0 + \boldsymbol{\beta}'\boldsymbol{x}_t + u_t$$

where

  - $\boldsymbol{x}_t = (x_{1t}, \ldots, x_{pt})'$ is a vector of p exogenous variables,
  - u_t is a zero-mean innovation,
  - $\boldsymbol{x}_t = [\boldsymbol{x}_t(S)', \boldsymbol{x}_t(S^c)']'$, $\boldsymbol{x}_t(S) \in \mathbb{R}^s$ is the vector of **relevant** variables and $\boldsymbol{x}_t(S^c) \in \mathbb{R}^{p-s}$ is the vector of **irrelevant** ones. $\boldsymbol{\beta} = [\boldsymbol{\beta}'_S, \boldsymbol{\beta}'_{S^c}]'$.
  - $p \equiv p(T)$ and $s \equiv s(T)$. T is the sample size.

- Goals:
  1. Select the right set of variables: $\widehat{\boldsymbol{\beta}}_S \neq \mathbf{0}$ and $\widehat{\boldsymbol{\beta}}_{S^c} = \mathbf{0}$ (model selection).
  2. Estimate $\boldsymbol{\beta}_S$ as if the correct set of variables is known to the econometrician.

Penalized Least Squares

- A Penalized Least Squares estimator $\widehat{\boldsymbol{\beta}}$:

$$\boxed{\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} \sum_{t=1}^{T} (y_t - \boldsymbol{\beta}' \boldsymbol{z}_t)^2 + \sum_{j=1}^{p} \color{red}{p_\lambda(|\beta_j|)},}$$

where

Penalized Least Squares

- A Penalized Least Squares estimator $\widehat{\boldsymbol{\beta}}$:

$$\boxed{\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \sum_{j=1}^{p} {\color{red}p_\lambda(|\beta_j|)},}$$

where

– $\boldsymbol{z}_t = (1, \boldsymbol{x}_t')'$.

Penalized Least Squares

- A Penalized Least Squares estimator $\widehat{\boldsymbol{\beta}}$:

$$\boxed{\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|),}$$

where

- $\boldsymbol{z}_t = (1, \boldsymbol{x}_t')'$.
- $p_\lambda(|\beta_j|)$ is a non-negative penalty function indexed by the **regularization parameter** λ. (e.g., $p_\lambda(|\beta_j|) = \lambda|\beta_j|^2$, or $p_\lambda(|\beta_j|) = \lambda|\beta_j|$).

Penalized Least Squares

- A Penalized Least Squares estimator $\widehat{\boldsymbol{\beta}}$:

$$\boxed{\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|),}$$

where

  - $\boldsymbol{z}_t = (1, \boldsymbol{x}_t')'$.
  - $p_\lambda(|\beta_j|)$ is a non-negative penalty function indexed by the **regularization parameter** λ. (e.g., $p_\lambda(|\beta_j|) = \lambda|\beta_j|^2$, or $p_\lambda(|\beta_j|) = \lambda|\beta_j|$).

- λ controls the "number of parameters" in the model.

Penalized Least Squares

- A Penalized Least Squares estimator $\widehat{\boldsymbol{\beta}}$:

$$\boxed{\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} \sum_{t=1}^{T} (y_t - \boldsymbol{\beta}' \boldsymbol{z}_t)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|),}$$

  where
  - $\boldsymbol{z}_t = (1, \boldsymbol{x}_t')'$.
  - $p_\lambda(|\beta_j|)$ is a non-negative penalty function indexed by the **regularization parameter** λ. (e.g., $p_\lambda(|\beta_j|) = \lambda|\beta_j|^2$, or $p_\lambda(|\beta_j|) = \lambda|\beta_j|$).
- λ controls the "number of parameters" in the model.
- If $\lambda = \infty$ no variables enter the model, if $\lambda = 0$ it is just the OLS estimator.

Penalized Least Squares

- A Penalized Least Squares estimator $\widehat{\boldsymbol{\beta}}$:

$$\boxed{\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|),}$$

  where
  - $\boldsymbol{z}_t = (1, \boldsymbol{x}_t')'$.
  - $p_\lambda(|\beta_j|)$ is a non-negative penalty function indexed by the **regularization parameter** λ. (e.g., $p_\lambda(|\beta_j|) = \lambda|\beta_j|^2$, or $p_\lambda(|\beta_j|) = \lambda|\beta_j|$).
- λ controls the "number of parameters" in the model.
- If $\lambda = \infty$ no variables enter the model, if $\lambda = 0$ it is just the OLS estimator.
- Key assumption (for some methods): **sparsity**.

Sparse Models

- We say a model is **sparse** if the *true* parameter vector $\boldsymbol{\beta}$ is **sparse**, i.e., most elements in $\boldsymbol{\beta}$ are either zero or negligibly small (compared to the sample size).

Sparse Models

- We say a model is **sparse** if the *true* parameter vector $\boldsymbol{\beta}$ is **sparse**, i.e., most elements in $\boldsymbol{\beta}$ are either zero or negligibly small (compared to the sample size).
- In some cases (for example, linear models for the conditional mean) it is equivalent to say that the number of **relevant** variables is small compared to the number of **candidate** variables.

Sparse Models

- We say a model is **sparse** if the *true* parameter vector $\boldsymbol{\beta}$ is **sparse**, i.e., most elements in $\boldsymbol{\beta}$ are either zero or negligibly small (compared to the sample size).
- In some cases (for example, linear models for the conditional mean) it is equivalent to say that the number of **relevant** variables is small compared to the number of **candidate** variables.
- Sparse modeling has been successfully used to deal with high-dimensional models and is a crucial condition for identifiability.

The Ridge Regression

- The Ridge estimator is defined as follows:

$$\boxed{\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda\sum_{j=0}^{p}\beta_j^2}$$

The Ridge Regression

- The Ridge estimator is defined as follows:

$$\boxed{\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda\sum_{j=0}^{p}\beta_j^2}$$

- "Shrinks" towards zero parameters associated with redundant predictors (not exactly).

The Ridge Regression

- The Ridge estimator is defined as follows:

$$\boxed{\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} \sum_{t=1}^{T} (y_t - \boldsymbol{\beta}' \boldsymbol{z}_t)^2 + \lambda \sum_{j=0}^{p} \beta_j^2}$$

- "Shrinks" towards zero parameters associated with redundant predictors (not exactly).
- λ is a shrinkage parameter to be chosen;

The Ridge Regression

- The Ridge estimator is defined as follows:

$$\boxed{\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T} (y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda \sum_{j=0}^{p} \beta_j^2}$$

- "Shrinks" towards zero parameters associated with redundant predictors (not exactly).
- λ is a shrinkage parameter to be chosen;
- The Ridge solution is not sparse.

The Ridge Regression

- The Ridge estimator is defined as follows:

$$\boxed{\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda \sum_{j=0}^{p} \beta_j^2}$$

- "Shrinks" towards zero parameters associated with redundant predictors (not exactly).
- λ is a shrinkage parameter to be chosen;
- The Ridge solution is not sparse.
- The solution $\widehat{\boldsymbol{\beta}}_{Ridge}$ is easy to find as the problem remains quadratic in $\boldsymbol{\beta}$:

$$\boxed{\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \left(\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y}.}$$

The Ridge Regression

- The Ridge estimator is defined as follows:

$$\boxed{\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \underset{\boldsymbol{\beta} \in \mathcal{B}}{\arg\min} \sum_{t=1}^{T} (y_t - \boldsymbol{\beta}' \boldsymbol{z}_t)^2 + \lambda \sum_{j=0}^{p} \beta_j^2}$$

- "Shrinks" towards zero parameters associated with redundant predictors (not exactly).
- λ is a shrinkage parameter to be chosen;
- The Ridge solution is not sparse.
- The solution $\widehat{\boldsymbol{\beta}}_{Ridge}$ is easy to find as the problem remains quadratic in $\boldsymbol{\beta}$:

$$\boxed{\widehat{\boldsymbol{\beta}}_{Ridge}(\lambda) = \left(\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{Z}'\boldsymbol{y}.}$$

- Good for prediction but not for variable selection.

The LASSO - Tibshirani (JRRS B, 1996)

- Least Absolute Shrinkage and Selection Operator (LASSO):

$$\boxed{\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda \sum_{j=1}^{p} |\beta_j|}$$

The LASSO - Tibshirani (JRRS B, 1996)

- Least Absolute Shrinkage and Selection Operator (LASSO):

$$\boxed{\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda\sum_{j=1}^{p}|\beta_j|}$$

- "Shrinks" to zero parameters associated with redundant predictors.

The LASSO - Tibshirani (JRRS B, 1996)

- Least Absolute Shrinkage and Selection Operator (LASSO):

$$\boxed{\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \arg\min_{\boldsymbol{\beta} \in \mathcal{B}} \sum_{t=1}^{T} (y_t - \boldsymbol{\beta}' \boldsymbol{z}_t)^2 + \lambda \sum_{j=1}^{p} |\beta_j|}$$

- "Shrinks" to zero parameters associated with redundant predictors.
- The regularization path can be efficiently estimated.

The LASSO - Tibshirani (JRRS B, 1996)

- Least Absolute Shrinkage and Selection Operator (LASSO):

$$\boxed{\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T} (y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda \sum_{j=1}^{p} |\beta_j|}$$

- "Shrinks" to zero parameters associated with redundant predictors.
- The regularization path can be efficiently estimated.
- Can handle (many) more variables than observations ($p >> T$).

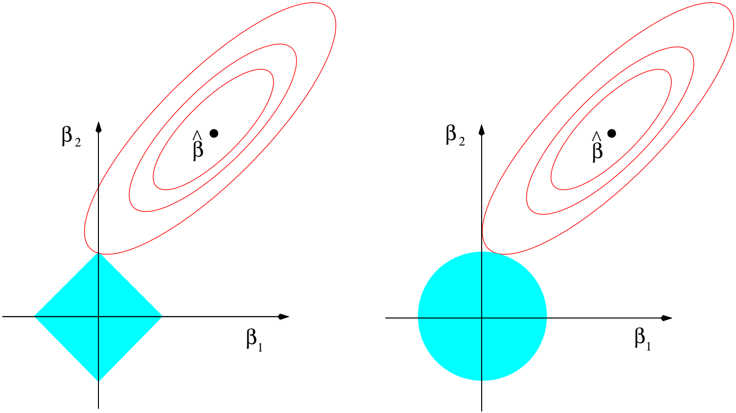
The LASSO - Tibshirani (JRRS B, 1996)

- Least Absolute Shrinkage and Selection Operator (LASSO):

$$\boxed{\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda \sum_{j=1}^{p} |\beta_j|}$$

- "Shrinks" to zero parameters associated with redundant predictors.
- The regularization path can be efficiently estimated.
- Can handle (many) more variables than observations ($p >> T$).
- Under some conditions can select the correct subset of relevant variables.

LASSO versus Ridge



Estimation Consistency

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0 \xrightarrow{p} \mathbf{0}, \text{ as } T \longrightarrow \infty.$$

Model Selection Consistency

$$\mathsf{P}\left(\left\{i : \widehat{\boldsymbol{\beta}} \neq \mathbf{0}\right\} = \left\{i : \boldsymbol{\beta}^0 \neq \mathbf{0}\right\}\right) \longrightarrow 1, \text{ as } T \longrightarrow \infty.$$

Sign Consistency

$$\mathsf{P}\left(\widehat{\boldsymbol{\beta}} \stackrel{s}{=} \boldsymbol{\beta}^0\right) \longrightarrow 1 \text{ as } T \longrightarrow \infty$$

where

$$\widehat{\boldsymbol{\beta}} \stackrel{s}{=} \boldsymbol{\beta}^0 \Longleftrightarrow \mathsf{sign}\left(\widehat{\boldsymbol{\beta}}\right) = \mathsf{sign}\left(\boldsymbol{\beta}^0\right)$$

The sign Function

The sign function is defined as

$$\boxed{\mathsf{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}}$$

Sign Consistency

Definitions

Strong Sign Consistency

The LASSO estimator is **strongly sign consistent** if $\exists \lambda_T = f(T)$ such that

$$\lim_{T\to\infty} \mathsf{P}\left(\widehat{\boldsymbol{\beta}}(\lambda_T) \stackrel{s}{=} \boldsymbol{\beta}^0\right) = 1$$

General Sign Consistency

The LASSO estimator is **general sign consistent** if

$$\lim_{T\to\infty} \mathsf{P}\left(\exists \lambda, \widehat{\boldsymbol{\beta}}(\lambda) \stackrel{s}{=} \boldsymbol{\beta}^0\right) = 1$$

- Strong sign consistency **implies** general sign consistency

General Sign Consistency versus Strong Sign Consistency

- **Strong Sign Consistency** implies one can use a pre-selected λ to achieve consistent model selection via the LASSO.

General Sign Consistency versus Strong Sign Consistency

- **Strong Sign Consistency** implies one can use a pre-selected λ to achieve consistent model selection via the LASSO.

- **General Sign Consistency** means for a random realization there exists a correct amount of regularization that selects the true model.

LASSO and Model Selection

Irrepresentable Condition

Strong Irrepresentable Condition

$\exists \eta > 0$ such that

$$\left| \widehat{\boldsymbol{\Sigma}}_{S^c S} \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \mathsf{sign}\left(\boldsymbol{\beta}_S^0\right) \right| \leq \mathbf{1} - \eta$$

Weak Irrepresentable Condition

$$\left| \widehat{\boldsymbol{\Sigma}}_{S^c S} \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \mathsf{sign}\left(\boldsymbol{\beta}_S^0\right) \right| < \mathbf{1}$$

- $\mathbf{1} \in \mathbb{R}^{(p-s)}$ is a vector of ones, and the inequality holds element-wise.
- The Irrepresentable Condition is a key condition for model selection consistency of the LASSO!

LASSO and Model Selection

Irrepresentable Condition

Strong Irrepresentable Condition

$\exists \eta > 0$ such that

$$\left| \widehat{\boldsymbol{\Sigma}}_{S^c S} \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \mathsf{sign}\left(\boldsymbol{\beta}_S^0\right) \right| \leq \mathbf{1} - \eta$$

Weak Irrepresentable Condition

$$\left| \widehat{\boldsymbol{\Sigma}}_{S^c S} \widehat{\boldsymbol{\Sigma}}_{SS}^{-1} \mathsf{sign}\left(\boldsymbol{\beta}_S^0\right) \right| < \mathbf{1}$$

- $\mathbf{1} \in \mathbb{R}^{(p-s)}$ is a vector of ones, and the inequality holds element-wise.
- The Irrepresentable Condition is a key condition for model selection consistency of the LASSO!
- This is a too strong condition!

The Adaptive LASSO - Zou (JASA, 2006)

- The Adaptive LASSO (adaLASSO) estimator is given by

$$\boxed{\widehat{\boldsymbol{\beta}}_{adaLASSO} = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|,}$$

where $w_1, \ldots, w_p$ are non-negative pre-defined weights.

The Adaptive LASSO - Zou (JASA, 2006)

- The Adaptive LASSO (adaLASSO) estimator is given by

$$\boxed{\widehat{\boldsymbol{\beta}}_{adaLASSO} = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|,}$$

  where $w_1, \ldots, w_p$ are non-negative pre-defined weights.
- Usually $w_j = |\tilde{\beta}_j|^{-\tau}$, for $\tau > 0$, where $\tilde{\beta}_j$ is an initial estimator (e.g., LASSO).

The Adaptive LASSO - Zou (JASA, 2006)

- The Adaptive LASSO (adaLASSO) estimator is given by

$$\boxed{\widehat{\boldsymbol{\beta}}_{adaLASSO} = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|,}$$

  where $w_1, \ldots, w_p$ are non-negative pre-defined weights.
- Usually $w_j = |\tilde{\beta}_j|^{-\tau}$, for $\tau > 0$, where $\tilde{\beta}_j$ is an initial estimator (e.g., LASSO).
- Provide consistent estimates for the non-zero parameters;

The Adaptive LASSO - Zou (JASA, 2006)

- The Adaptive LASSO (adaLASSO) estimator is given by

$$\boxed{\widehat{\boldsymbol{\beta}}_{adaLASSO} = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|,}$$

  where $w_1, \ldots, w_p$ are non-negative pre-defined weights.
- Usually $w_j = |\tilde{\beta}_j|^{-\tau}$, for $\tau > 0$, where $\tilde{\beta}_j$ is an initial estimator (e.g., LASSO).
- Provide consistent estimates for the non-zero parameters;
- Has the oracle property under some conditions.

The Adaptive LASSO - Zou (JASA, 2006)

- The Adaptive LASSO (adaLASSO) estimator is given by

$$\boxed{\widehat{\boldsymbol{\beta}}_{adaLASSO} = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda \sum_{j=1}^{p} w_j|\beta_j|,}$$

  where $w_1, \dots, w_p$ are non-negative pre-defined weights.
- Usually $w_j = |\tilde{\beta}_j|^{-\tau}$, for $\tau > 0$, where $\tilde{\beta}_j$ is an initial estimator (e.g., LASSO).
- Provide consistent estimates for the non-zero parameters;
- Has the oracle property under some conditions.
- Theoretical results in general time-series framework: Medeiros and Mendes (JoE, 2016)

The Elastic Net Estimator

- The Naïve Elastic Net estimator is defined as

$$\boxed{\widehat{\boldsymbol{\beta}}(\text{naïve}) = \arg\min_{\boldsymbol{\beta}\in\in\mathbb{R}^p} \sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda_2 \sum_{j=1}^{p} \beta_i^2 + \lambda_1 \sum_{j=1}^{p} |\beta_i|.}$$

The Elastic Net Estimator

- The Naïve Elastic Net estimator is defined as

$$\boxed{\widehat{\boldsymbol{\beta}}(\text{naïve}) = \underset{\boldsymbol{\beta} \in \in \mathbb{R}^p}{\arg\min} \sum_{t=1}^{T} (y_t - \boldsymbol{\beta}' \boldsymbol{z}_t)^2 + \lambda_2 \sum_{j=1}^{p} \beta_i^2 + \lambda_1 \sum_{j=1}^{p} |\beta_i|.}$$

- The Elastic Net estimator is given by

$$\boxed{\widehat{\boldsymbol{\beta}} = (1 + \lambda_2) \widehat{\boldsymbol{\beta}}(\text{naïve}).}$$

The Elastic Net Estimator

- The Naïve Elastic Net estimator is defined as

$$\boxed{\widehat{\boldsymbol{\beta}}(\text{naïve}) = \underset{\boldsymbol{\beta}\in\in\mathbb{R}^p}{\arg\min} \sum_{t=1}^{T} (y_t - \boldsymbol{\beta}'\boldsymbol{z}_t)^2 + \lambda_2 \sum_{j=1}^{p} \beta_i^2 + \lambda_1 \sum_{j=1}^{p} |\beta_i|.}$$

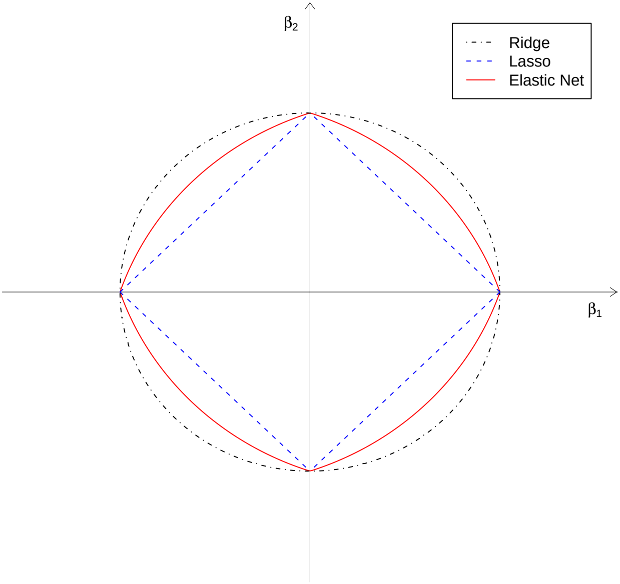
- The Elastic Net estimator is given by

$$\boxed{\widehat{\boldsymbol{\beta}} = (1+\lambda_2)\widehat{\boldsymbol{\beta}}(\text{naïve}).}$$

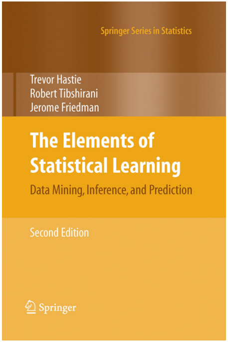
- The naïve EL-Net estimator selects the same model as the EL-Net version.

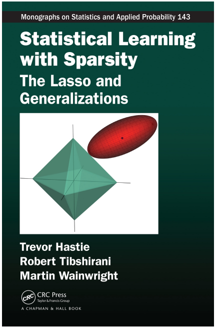
The Elastic Net Estimator

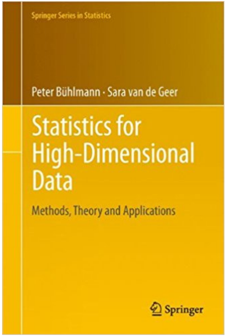
The Geometry of the Elastic Net



To Learn More about Shrinkage







Empirical Example: Forecasting Large Dimensional Realized Covariance Matrices

Callot, Laurent, Anders B. Kock and Marcelo C. Medeiros (2017). *Modeling and Forecasting Large Realized Covariance Matrices and Portfolio Choice*. **Journal of Applied Econometrics**, 32, 140-158.

Dataset

- 30 stocks from the Dow Jones index from 2006 to 2012 with a total of 1474 daily observations.
- Daily realized covariances are constructed from 5 minutes returns by the method of Lunde, Shephard, Sheppard (2013).
- The stocks can be classified in 8 broad categories.

Basic Materials	Technology	Consumer Cyclical	Consumer Non-cyclical
2	4	3	7
Energy	Financial	Industrial	Communication
2	3	5	4

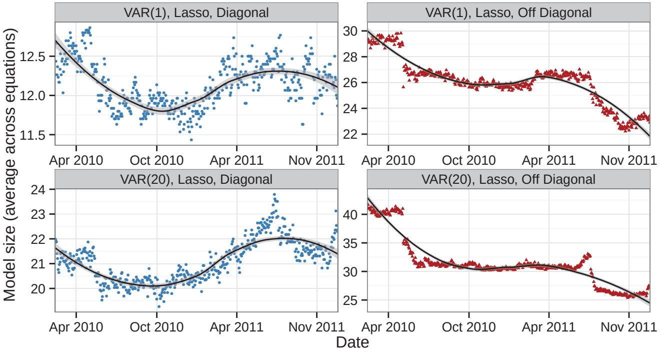
Results: Sectors

		Variance Equations							
Lagged variance	Basic Materials	**0.75**	0.40	0.14	0.52	0.23	0.35	0.57	0.39
	Consumer, Non-cyclical	0.17	0.48	0.37	0.37	0.24	0.20	0.26	0.32
	Financial	0.00	0.42	**0.99**	0.24	0.64	0.20	0.12	0.48
	Communications	0.32	0.23	0.10	**0.57**	0.19	0.14	0.27	0.19
	Industrial	0.00	0.19	0.28	0.16	**1.00**	0.08	0.07	0.18
	Energy	0.58	0.45	0.46	0.33	0.02	**1.00**	0.38	0.55
	Technology	0.34	0.19	0.09	0.24	0.02	0.05	**0.63**	0.12
	Consumer, Cyclical	0.34	**0.54**	0.35	0.29	0.30	0.20	0.31	**0.70**
Lagged covariance	Basic Materials	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.01
	Consumer, Non-cyclical	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
	Financial	0.00	0.01	0.02	0.00	0.00	0.00	0.01	0.00
	Communications	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Industrial	0.02	0.01	0.01	0.02	0.03	0.00	0.03	0.02
	Energy	0.01	0.03	0.01	0.03	0.01	0.01	0.02	0.03
	Technology	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00
	Consumer, Cyclical	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Results: Sectors

		Covariance Equations							
Lagged variance	Basic Materials	**0.81**	0.27	0.10	0.24	0.24	**0.93**	0.26	0.29
	Consumer, Non-cyclical	0.48	**0.71**	0.56	0.32	0.28	0.35	0.36	0.41
	Financial	0.13	0.25	**0.64**	0.16	0.06	0.32	0.15	0.17
	Communications	0.62	0.57	0.54	**0.65**	**0.51**	0.70	0.61	0.58
	Industrial	0.13	0.13	0.18	0.18	0.34	0.08	0.10	0.07
	Energy	0.12	0.08	0.21	0.06	0.03	0.56	0.11	0.12
	Technology	0.74	0.49	0.51	0.52	0.43	0.34	**0.82**	0.53
	Consumer, Cyclical	0.14	0.52	0.55	0.37	0.37	0.51	0.29	**0.90**
Lagged covariance	Basic Materials	0.09	0.12	0.12	0.11	0.09	0.12	0.09	0.11
	Consumer, Non-cyclical	0.04	0.05	0.05	0.04	0.03	0.05	0.04	0.03
	Financial	0.11	0.14	0.18	0.11	0.07	0.10	0.10	0.10
	Communications	0.07	0.09	0.09	0.11	0.04	0.10	0.08	0.08
	Industrial	0.17	0.14	0.15	0.14	0.24	0.11	0.14	0.15
	Energy	0.16	0.16	0.16	0.15	0.10	0.24	0.15	0.16
	Technology	0.08	0.07	0.09	0.08	0.05	0.07	0.08	0.06
	Consumer, Cyclical	0.05	0.05	0.05	0.06	0.04	0.08	0.05	0.06

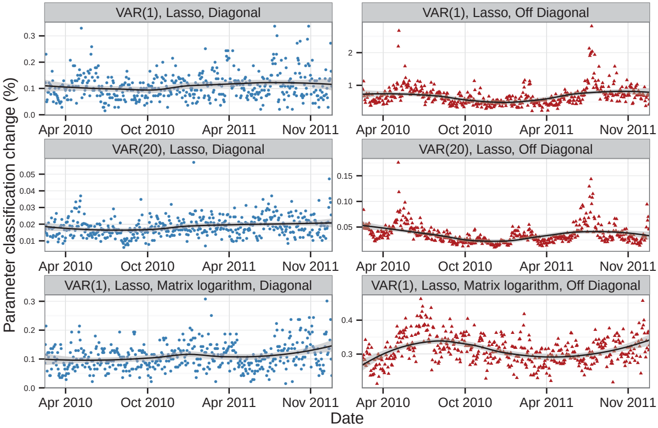
Results: Average Equation Size



- Diagonal equations more stable than off-diagonal ones.
- Diagonal equations are smaller.
- Flash Crash: May 6th 2010.

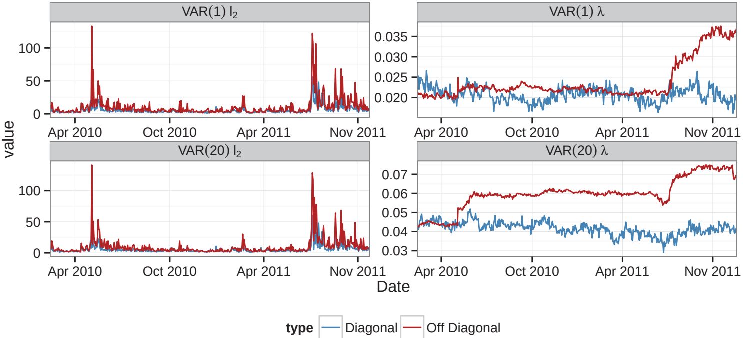
Results: Parameter Stability

Fraction of parameters that change from being zero to non-zero or vice versa in two consecutive periods

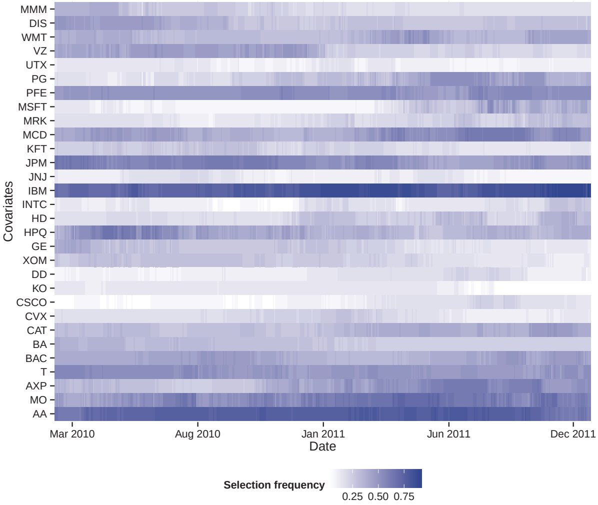


Results: Forecast Error and Penalty Parameter

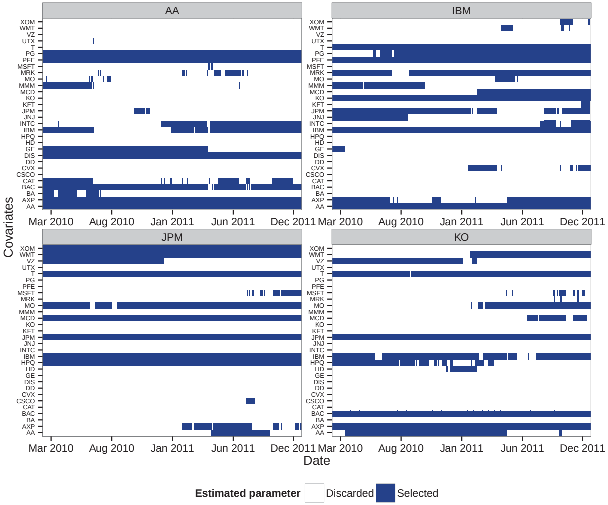
ℓ_2-norm of the 1-step ahead forecast error (left panel) and average penalty parameter (right panel) selected by BIC.



Results: Selection Frequency – VAR(1) LASSO

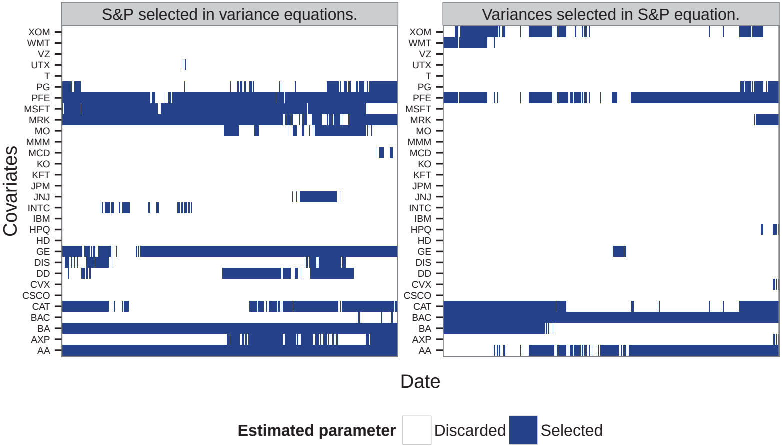


Results: Selection – VAR(1) LASSO



Results: Common Factor – VAR (1) LASSO

Lagged variance of the S&P selected in the variance equations of the Dow Jones stocks (left panel) and lagged variances of the Dow Jones stocks selected in the equation of the variance of the S&P 500 (right panel).



Forecasting Results

		AMedAFE			AMaxAFE			ℓ_2		
Model	h	A	D	O	A	D	O	A	D	O
No-Change	1	0.33	0.57	0.33	3.53	3.53	1.47	11.22	5.98	9.22
Censored	5	0.46	0.79	0.45	4.51	4.51	1.91	15.02	7.89	12.41
	20	0.58	0.98	0.57	5.12	5.12	2.22	18.05	9.25	15.17
DCC	1	0.56	0.95	0.55	8.40	8.36	4.28	22.37	12.40	18.17
EWMA($\lambda = 0.96$)	1	0.88	1.08	0.88	8.07	8.03	4.55	28.89	12.55	25.78
VAR(1), Lasso	1	0.37	0.61	0.37	3.34	3.32	1.72	11.98	5.93	10.21
	5	0.44	0.73	0.43	3.77	3.64	2.25	14.25	6.82	12.27
	20	0.69	0.96	0.68	4.37	4.03	3.16	19.98	8.11	18.07
VAR(1), Lasso	1	0.34	0.55	0.33	3.08	3.04	1.76	11.26	5.4	9.72
Post Lasso OLS	5	0.45	0.73	0.44	3.8	3.68	2.23	14.39	6.87	12.36
	20	0.61	0.93	0.6	4.34	4.09	2.94	18.55	8.06	16.43
VAR(1), adaptive Lasso	1	0.37	0.62	0.37	3.46	3.44	1.81	12.21	6.07	10.4
Initial estimator: Lasso	5	0.44	0.74	0.44	3.88	3.78	2.32	14.49	6.93	12.52
	20	0.62	0.98	0.61	4.45	4.18	3.13	19.44	8.38	17.3
VAR(1), Lasso	1	0.35	0.58	0.35	3.25	3.25	1.42	11.31	5.76	9.48
Log-matrix transform	5	0.42	0.73	0.41	3.58	3.58	1.62	13.26	6.65	11.2
	20	0.48	0.94	0.47	4.02	4.02	1.81	15.27	8.04	12.64
VAR(1), Lasso	1	0.37	0.61	0.37	3.34	3.32	1.77	12.44	5.93	10.22
Including S&P 500	5	0.44	0.74	0.43	3.79	3.65	2.33	14.77	6.84	12.31
	20	0.68	0.95	0.67	4.37	4.02	3.16	20.46	8.08	17.96

Forecasting Results

		AMedAFE			AMaxAFE			ℓ_2		
Model	h	A	D	O	A	D	O	A	D	O
No-Change	1	0.33	0.57	0.33	3.53	3.53	1.47	11.22	5.98	9.22
Censored	5	0.46	0.79	0.45	4.51	4.51	1.91	15.02	7.89	12.41
	20	0.58	0.98	0.57	5.12	5.12	2.22	18.05	9.25	15.17
DCC	1	0.56	0.95	0.55	8.40	8.36	4.28	22.37	12.40	18.17
EWMA($\lambda = 0.96$)	1	0.88	1.08	0.88	8.07	8.03	4.55	28.89	12.55	25.78
VAR(20), Lasso	1	0.35	0.57	0.35	3.19	3.16	1.62	11.35	5.59	9.66
	5	0.41	0.65	0.4	3.54	3.46	2.01	13.09	6.28	11.25
	20	0.54	0.84	0.53	4.03	3.87	2.56	16.29	7.44	14.3
VAR(20), Lasso	1	0.33	0.52	0.32	3.01	2.92	1.76	10.88	5.09	9.44
Post Lasso OLS	5	0.42	0.66	0.41	3.56	3.48	2.1	13.43	6.31	11.65
	20	0.49	0.79	0.47	4.02	3.9	2.38	15.29	7.27	13.25
VAR(20), adaptive Lasso	1	0.36	0.59	0.35	3.45	3.44	1.61	11.76	5.98	9.89
Initial estimator: Lasso	5	0.43	0.69	0.42	3.75	3.72	2.01	13.62	6.66	11.65
	20	0.58	0.93	0.57	4.16	4.04	2.68	17.49	8.03	15.33
VAR(20), Lasso	1	0.36	0.57	0.35	3.16	3.16	1.39	11.22	5.59	9.49
Log-matrix transform	5	0.4	0.66	0.39	3.42	3.42	1.54	12.53	6.22	10.63
	20	0.46	0.84	0.45	3.81	3.8	1.73	14.37	7.36	12.06
VAR(20), Lasso	1	0.35	0.57	0.35	3.19	3.16	1.64	11.78	5.59	9.67
Including S&P 500	5	0.41	0.65	0.4	3.54	3.46	2.01	13.55	6.28	11.26
	20	0.54	0.84	0.53	4	3.85	2.54	16.83	7.4	14.38

Portfolio Results

The investor's problem at $t = t_0, \ldots, T-1$ is to select a vector of weights for period $t+1$ based solely on information up to time t:

$$
\begin{aligned}
\widehat{\omega}_{t+1} = \arg\min_{\omega_{t+1}} & \, \omega'_{t+1}\widehat{\Sigma}_{t+1}\omega_{t+1} \\
\text{s.t.} \quad & \omega'_{t+1}\widehat{\mu}_{t+1} = \mu_{\textsf{target}} \\
& \sum_{i=1}^{n} \omega_{it+1} = 1 \\
& \sum_{i=1}^{n} |\omega_{it+1}| \mathsf{I}(\omega_{it} < 0) \leq 0.30 \\
& |\omega_{it+1}| \leq 0.20,
\end{aligned}
$$

where ω_{t+1} is an $n \times 1$ vector of portfolio weights, $\mu_{\textsf{target}}$ is the target expected rate of return from t to $t+1$ and $\mathsf{I}(\cdot)$ is an indicator function.

Portfolio Results

Model	VAR(1)				VAR(20)				No-Change	DCC	EWMA
Estimator: Statistic	Lasso	Post Lasso OLS	adaLasso Init: Lasso	Lasso (Log Mat)	Lasso	Post Lasso OLS	adaLasso Init: Lasso	Lasso (Log Mat)	Censored		
Average weight	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Max weight	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
Min weight	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20	-0.20
Average leverage	0.28	0.28	0.28	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29
Proportion of leverage	0.22	0.23	0.22	0.27	0.22	0.22	0.22	0.28	0.24	0.24	0.23
Average turnover	0.02	0.03	0.03	0.02	0.02	0.02	0.02	0.01	0.03	0.01	0.01
Average return ($\times 10^{-4}$)	2.58	2.72	3.58	5.89	2.85	2.96	2.64	6.27	1.99	0.40	0.14
Accumulated return	10.07	10.71	15.16	27.77	11.45	11.93	10.34	30.02	7.13	-0.48	-1.60
Standard deviation ($\times 10^2$)	0.97	0.98	0.98	1.00	0.97	0.99	0.97	0.99	0.97	1.00	0.99
Sharpe ratio ($\times 10^2$)	2.66	2.77	3.65	5.87	2.95	3.01	2.70	6.29	2.04	0.40	0.14
Diversification ratio	1.46	1.46	1.47	1.43	1.46	1.46	1.44	1.43	1.48	1.43	1.43
Economic Value $\gamma = 1$											
No-Change (censored)	1.50	1.83	4.08	10.25	2.21	2.45	1.64	11.32	–	–	–
DCC	5.73	6.07	8.41	14.84	6.46	6.72	5.87	15.95	–	–	–
EWMA	6.40	6.74	9.09	15.56	7.13	7.39	6.54	16.68	–	–	–
Economic Value $\gamma = 5$											
No-Change (censored)	1.54	1.77	4.00	9.92	2.27	2.33	1.64	11.06	–	–	–
DCC	6.13	6.37	8.71	14.89	6.90	6.96	6.24	16.08	–	–	–
EWMA	6.68	6.92	9.27	15.48	7.45	7.51	6.79	16.68	–	–	–
Economic Value $\gamma = 10$											
No-Change (censored)	1.58	1.68	3.91	9.50	2.35	2.17	1.63	10.74	–	–	–
DCC	6.64	6.75	9.08	14.95	7.45	7.26	6.69	16.25	–	–	–
EWMA	7.04	7.15	9.49	15.38	7.85	7.66	7.09	16.69	–	–	–

Empirical Example: Forecasting Even Larger Realized Covariance Matrices

Brito, Diego, Marcelo C. Medeiros and Ruy M. Ribeiro (2018). *Forecasting Large Realized Covariance Matrices: The Benefits of Factor Models and Shrinkage*. Working paper available at SSRN id 3163668.

The Setup

Curse of Dimensionality

- RC matrices are highly persistent over time, which suggests an AR model of large order p (usually $p > 20$).
- $\boldsymbol{\Sigma}_t$: $n \times n$ realized covariance matrix.
- $\boldsymbol{y}_t = \mathsf{vech}(\boldsymbol{\Sigma}_t)$, such that

$$\boxed{\boldsymbol{y}_t = \boldsymbol{\omega} + \sum_{i=1}^{p} \boldsymbol{\Phi}_i y_{t-i} + \boldsymbol{\epsilon}_t,\ t = 1, \ldots, T,}$$

  where:
  - $\boldsymbol{\Phi}_i,\ i = 1, \ldots, p$ are the $q \times q$ matrices with $q = n(n+1)/2$;
  - $\boldsymbol{\omega}$ is a $q \times 1$ vector of intercepts.
- $n(n+1)/2$ equations with a total of $n(n+1)(p+1)/2$ parameters.

The Setup

Factor Structure

- Excess return on any asset i, $r_{i,t}$:

$$
\begin{aligned}
r^e_{i,t} &= \beta_{i1,t} f_{1,t} + \cdots + \beta_{iK,t} f_{K,t} + \varepsilon_{i,t} = \boldsymbol{\beta}'_{i,t} \boldsymbol{f}_t + \varepsilon_{i,t}, \\
\boldsymbol{r}^e_t &= \boldsymbol{B}'_t \boldsymbol{f}_t + \boldsymbol{\varepsilon}_t,
\end{aligned}
$$

  where:
  - $f_{1,t}, \cdots, f_{K,t}$ are the excess returns of K factors;
  - $\beta_{ik,t}$, $k = 1, \ldots, K$, are factor loadings for asset i;
  - $\varepsilon_{i,t}$ is the idiosyncratic error term.

- Factors are linear combinations of returns: long-short stock portfolios where stocks are sorted on firm characteristics:

$$\boldsymbol{f}_t = \boldsymbol{W}_t \boldsymbol{r}^e_t \qquad \boldsymbol{W}_t \text{ is known}$$

- Loadings are time-varying and are given as:

$$\boldsymbol{B}_t = (\boldsymbol{\Sigma}_{f,t})^{-1} \boldsymbol{W}'_t \boldsymbol{\Sigma}_t$$

The Setup

Covariance Decomposition

- Under the assumption $\mathbb{E}(\boldsymbol{\varepsilon}_t|\boldsymbol{f}_t) = \mathbf{0}$, we have

$$\boxed{\boldsymbol{\Sigma}_t = \mathrm{cov}(\boldsymbol{B}_t'\boldsymbol{f}_t) + \mathrm{cov}(\boldsymbol{\varepsilon}_t) = \boldsymbol{B}_t'\boldsymbol{\Sigma}_{f,t}\boldsymbol{B}_t + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t}.}$$

- By linearity:

$$\boldsymbol{\Sigma}_{f,t} = \mathrm{cov}(\boldsymbol{f}_t) = \mathrm{cov}(\boldsymbol{W}_t'\boldsymbol{r}_t) = \boldsymbol{W}_t'\boldsymbol{\Sigma}_t\boldsymbol{W}_{\boldsymbol{t}}.$$

- Therefore,

$$\boxed{\widehat{\boldsymbol{\Sigma}}_{t+1|t} = \widehat{\boldsymbol{B}}'_{t+1|t}\widehat{\boldsymbol{\Sigma}}_{f,t+1|t}\widehat{\boldsymbol{B}}_{t+1|t} + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon},t+1|t}.}$$

Forecasting Methodology: $\widehat{\boldsymbol{\Sigma}}_{t+1|t} = \widehat{\boldsymbol{B}}'_{t+1|t}\widehat{\boldsymbol{\Sigma}}_{f,t+1|t}\widehat{\boldsymbol{B}}_{t+1|t} + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon},t+1|t}$

Realized Factor Covariance Matrices

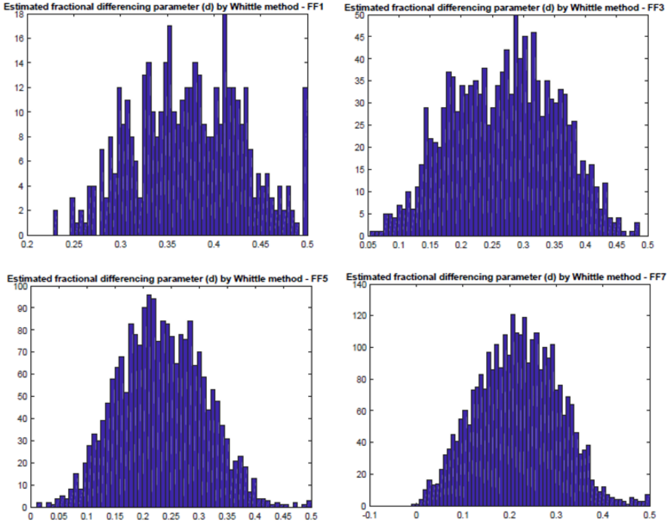
- Vector HAR model for $\boldsymbol{y}_{f,t} = \mathsf{vech}[\mathrm{logM}(\boldsymbol{\Sigma}_{f,t})]$:

$$\boxed{\boldsymbol{y}_{f,t} = \boldsymbol{\omega} + \boldsymbol{\Phi}_{\mathsf{day}}\boldsymbol{y}^{\mathsf{day}}_{f,t-1} + \boldsymbol{\Phi}_{\mathsf{week}}\boldsymbol{y}^{\mathsf{week}}_{f,t-1} + \boldsymbol{\Phi}_{\mathsf{month}}\boldsymbol{y}^{\mathsf{month}}_{f,t-1} + \boldsymbol{\epsilon}_t,}$$

  where:
  - $\boldsymbol{y}^{\mathsf{day}}_{f,t} = \mathsf{vech}(\boldsymbol{\Sigma}^{\mathsf{day}}_{f,t})$; $\boldsymbol{y}^{\mathsf{week}}_{f,t} = \mathsf{vech}(\boldsymbol{\Sigma}^{\mathsf{week}}_{f,t})$; $\boldsymbol{y}^{\mathsf{month}}_{f,t} = \mathsf{vech}(\boldsymbol{\Sigma}^{\mathsf{month}}_{f,t})$;
  - $\boldsymbol{\Sigma}^{\mathsf{day}}_{f,t} = \mathrm{logM}(\boldsymbol{\Sigma}_{f,t})$;
  - $\boldsymbol{\Sigma}^{\mathsf{week}}_{f,t} = \frac{1}{5}[\mathrm{logM}(\boldsymbol{\Sigma}_{f,t}) + \cdots + \mathrm{logM}(\boldsymbol{\Sigma}_{f,t-4})]$; and
  - $\boldsymbol{\Sigma}^{\mathsf{month}}_{f,t} = \frac{1}{22}[\mathrm{logM}(\boldsymbol{\Sigma}_{f,t}) + \cdots + \mathrm{logM}(\boldsymbol{\Sigma}_{f,t-21})]$.
- Estimation via LASSO/adaLASSO
- Penalty parameter is set with the BIC
- The inverse LASSO estimates (in absolute value) are used as weights for the adaLASSO

Forecasting Methodology: $\widehat{\boldsymbol{\Sigma}}_{t+1|t} = \widehat{\boldsymbol{B}}'_{t+1|t}\widehat{\boldsymbol{\Sigma}}_{f,t+1|t}\widehat{\boldsymbol{B}}_{t+1|t} + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon},t+1|t}$

Loadings



- Loading dynamics modeled as a HAR model:

$$\boxed{\beta_{k,i,t} = \omega + \phi_{\mathsf{day}}\beta^{\mathsf{day}}_{k,i,t-1} + \phi_{\mathsf{week}}\beta^{\mathsf{week}}_{k,i,t-1} + \phi_{\mathsf{month}}\beta^{\mathsf{month}}_{k,i,t-1} + \epsilon_{k,i,t},}$$

where $\beta_{k,i,t}$ is the (k,i) element of $\boldsymbol{B}_t$, i.e., the loading of stock i on factor k at date t.

- Coefficients estimated by OLS.

- No need for LASSO here.

Forecasting Methodology: $\widehat{\boldsymbol{\Sigma}}_{t+1|t} = \widehat{\boldsymbol{B}}'_{t+1|t}\widehat{\boldsymbol{\Sigma}}_{f,t+1|t}\widehat{\boldsymbol{B}}_{t+1|t} + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon},t+1|t}$

Residual Covariance

- Forecasting $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t}$ is still subject to the curse of dimensionality
- We assume that $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t}$ is **block-diagonal** where blocks are defined by industry classification.
- Furthermore, we assume that the dynamics of each block depends **only** on the elements of the same block at $t-1$
- Finally, **past covariances are not used as regressors** (Callot, Kock, and Medeiros, 2017)

Forecasting Methodology: $\widehat{\boldsymbol{\Sigma}}_{t+1|t} = \widehat{\boldsymbol{B}}'_{t+1|t}\widehat{\boldsymbol{\Sigma}}_{f,t+1|t}\widehat{\boldsymbol{B}}_{t+1|t} + \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\varepsilon},t+1|t}$

Residual Covariance

- S sectors:

$$\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon},t} = \begin{pmatrix} \boldsymbol{\Sigma}^1_{\boldsymbol{\varepsilon},t} & & \\ & \ddots & \\ & & \boldsymbol{\Sigma}^S_{\boldsymbol{\epsilon},t} \end{pmatrix}.$$

- The dynamics for $\boldsymbol{y}^s_{\boldsymbol{\varepsilon},t} = \mathsf{vech}[\mathrm{logM}(\boldsymbol{\Sigma}^s_{\boldsymbol{\varepsilon},t})], s \in \{1, 2, \ldots, S\}$:

$$\boxed{\boldsymbol{y}^s_{\boldsymbol{\varepsilon},t} = \boldsymbol{\omega}^s_{\epsilon} + \boldsymbol{\Phi}^s \boldsymbol{\Lambda}^s_{\boldsymbol{\varepsilon},t-1} + \boldsymbol{u}^s_{\boldsymbol{\varepsilon},t},}$$

where $\boldsymbol{\Lambda}^s_{\boldsymbol{\varepsilon},t-1} = \mathrm{diag}[\mathrm{logM}(\boldsymbol{\Sigma}^s_{\boldsymbol{\varepsilon},t-1})]$.

- LASSO/adaLASSO estimation equation by equation.

Data

Realized Covariance Matrices

- The data consists of daily realized covariance matrices of returns for constituents of the S&P 500 index
- We consider companies that remained in the index and had balance sheet data for the full sample period, totaling 430 stocks
- These matrices were constructed from 5-minute returns by composite realized kernel (Lunde et al, 2016 JBES)
- Sample period: January 2006 - December 2011 (1495 days). Estimation windows with 1,000 observations.
- Data cleaning: merges and splits.

Data

Factors and Sector Classification

- **6 factors + market are considered**: Size (SMB), Value (HML), Gross Profitability, Investment, Asset Growth and Accruals (CRSP/Compustat database)
- **4 different combinations**: 1F(Market), 3F(1F + Size and Value), 5F(3F + Gross Profitability and Investment), and 7F(5F + Asset Growth and Accruals)
- **Standard Industrial Classification (SIC)**: 10 sectors

Data

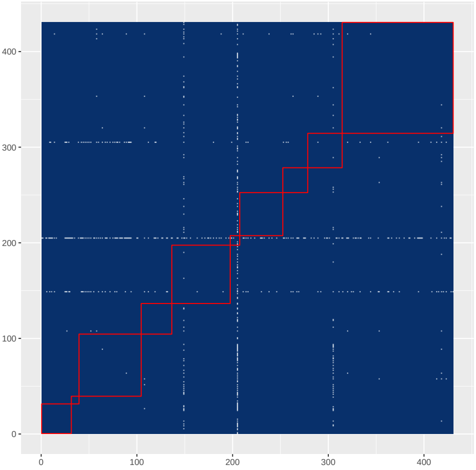
Number of Stocks per Sector

Sector	Number of Stocks
Consumer Non-Durables	31
Consumer Durables	8
Manufacturing	65
Oil, Gas, and Coal Extraction	32
Business Equipment	61
Telecommunications	10
Wholesale and Retail	45
Health Care, Medical Equipments, and Drugs	26
Utilities	36
Others	116

Results

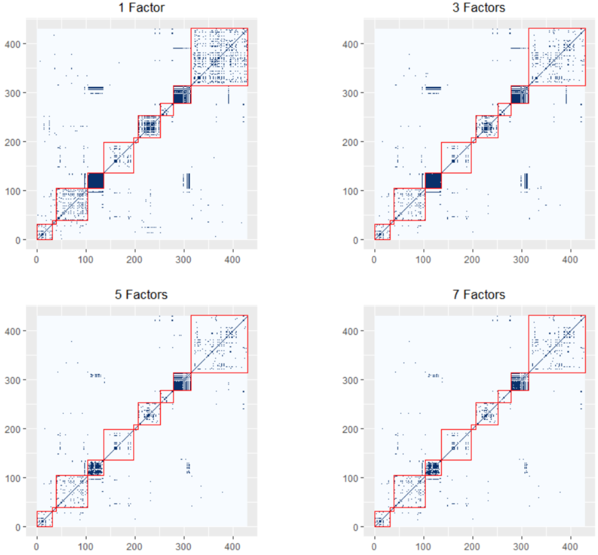
Covariance Structure

- The blue dots represent the correlations larger than 0.15 in absolute value in at least 1/3 of the sample days.
- Red squares represent the groups defined by SIC.



Results

Factor Decomposition and Residual Covariance



Forecasting Results

Forecast Precision for Factor Covariance Matrices

- ℓ_2 represents the average ℓ_2-forecast error over the 473 out-of-sample days, that is,

$$\text{average } \ell_2\text{-forecast error} = \frac{1}{T_2 - T_1 + 1} \sum_{T=T_1}^{T=T_2} ||\widehat{\boldsymbol{\epsilon}}_{T+1}||.$$

- $\ell_2/\ell_{2,RW}$ represents the ratio of the ℓ_2-forecast error for other methods to the random walk value.

	ℓ_2	ℓ_2 / $\ell_{2,RW}$	
Model	Random Walk	FHAR	FHAR, Log-matrix
1F	0.40	0.96 (0.96)	**0.92 (0.92)**
3F	0.44	0.98 (0.97)	**0.90 (0.90)**
5F	0.51	0.95 (0.95)	**0.89 (0.89)**
7F	0.62	0.99 (1.04)	**0.86 (0.87)**

Forecasting Results

Forecast Precision for Complete Covariance Matrices

Model (Benchmarks)	$\ell_2/\ell_{2,RW}$	VHAR (Log-matrix)	$\ell_2/\ell_{2,RW}$
RW	1.00	1F, LASSO	**0.86**
EWMA (Returns)	6.93	3F, LASSO	**0.85**
BEKK-NL	1.71	5F, LASSO	**0.85**
DCC-NL	1.71	7F, LASSO	**0.85**
Block 1F	0.97	1F, adaLASSO	0.86
Block 3F	0.97	3F, adaLASSO	0.85
Block 5F	0.97	5F, adaLASSO	0.85
Block 7F	0.97	7F, adaLASSO	0.85
Random Walk (RW) $\ell_{2,RW}$	341.57		

Portfolio Results

Statistics for Daily Portfolios - Global Minimum Variance

- Consider the problem of an investor at time $t = t_0, \ldots, T-1$ who wishes to construct a minimum variance portfolio to be held in time $t+1$.
- The optimization problem consists of choosing a vector of weights $\widehat{\boldsymbol{w}}_{t+1}$:

$$\boxed{\begin{aligned} \widehat{\boldsymbol{w}}_{t+1} = \underset{\boldsymbol{w}_{t+1}}{\arg\min} \quad & \boldsymbol{w}_{t+1}' \widehat{\boldsymbol{\Sigma}}_{t+1} \boldsymbol{w}_{t+1} \\ \text{subject to} \quad & \boldsymbol{w}_{t+1}' \mathbf{1} = 1. \end{aligned}}$$

Portfolio Results

Statistics for Daily Portfolios - Global Minimum Variance

	RW	Block 1F	Block 3F	Block 5F	Block 7F	EWMA	BEKK-NL	DCC - NL
Standard Deviation (%)	12.07	8.21	8.29	8.25	8.25	14.62	9.41	10.65
Lower Partial SD (%)	12.82	8.79	8.94	8.73	8.83	14.90	9.63	11.31
Avg. Gross Leverage	5.94	3.08	3.14	3.14	3.19	12.55	5.09	4.11
Prop. of Leverage (%)	44.30	44.40	44.22	44.10	44.11	49.17	45.11	51.73
Avg. Turnover (%)	1.80	0.75	0.78	0.78	0.80	0.27	0.11	0.21
Avg. Excess Return (%)	14.20	12.72	14.46	15.37	14.95	3.42	17.98	17.46
Cumulative Return (%)	29.04	26.42	30.59	32.86	31.82	4.74	39.27	37.58
Sharpe Ratio	1.18	1.55	1.74	1.86	1.81	0.23	1.91	1.64

	1 Factor VHAR (Log matrix)		3 Factors VHAR (Log matrix)		5 Factors VHAR (Log matrix)		7 Factors VHAR (Log matrix)	
	LASSO	aLASSO	LASSO	aLASSO	LASSO	aLASSO	LASSO	aLASSO
Standard Deviation (%)	8.46	8.42	8.37	8.32	8.29	8.25	8.12	8.09
Lower Partial SD (%)	8.86	8.81	8.78	8.68	8.57	8.53	8.52	8.51
Avg. Gross Leverage	2.66	2.67	2.80	2.80	2.82	2.82	2.93	2.93
Prop. of Leverage (%)	45.89	46.01	44.88	45.03	44.89	45.12	45.26	45.50
Avg. Turnover (%)	0.20	0.22	0.20	0.22	0.19	0.21	0.20	0.22
Avg. Excess Return (%)	15.24	15.18	17.69	17.45	18.93	18.61	18.09	17.85
Cumulative Return (%)	32.49	32.35	38.74	38.13	42.01	41.19	39.85	39.21
Sharpe Ratio	1.80	1.80	2.11	2.10	2.28	2.26	2.23	2.21

Portfolio Results

Statistics for Daily Portfolios - Restricted Minimum Variance

- Maximum leverage equal to 30% (in some sense, consistent with a 130-30 fund concept in the mutual fund industry).
- Maximum weights on individual stocks: 20% (in absolute value).
- The problem for an investor at time $t = t_0, \ldots, T-1$ is then given by

$$\boxed{\begin{gathered} \widehat{\boldsymbol{w}}_{t+1} = \underset{\boldsymbol{w}_{t+1}}{\arg\min} \quad \boldsymbol{w}'_{t+1} \widehat{\boldsymbol{\Sigma}}_{t+1} w_{t+1} \\ \text{subject to} \quad \boldsymbol{w}'_{t+1} \mathbf{1} = 1, \\ \sum_{i=1}^{N} |w_{it+1}| I(w_{it} < 0) \leq 0.30 \quad \text{and} \quad |w_{it+1}| \leq 0.20. \end{gathered}}$$

Portfolio Results

Statistics for Daily Portfolios - Restricted Minimum Variance

	RW	Block 1F	Block 3F	Block 5F	Block 7F	EWMA	BEKK-NL	DCC - NL
Standard Deviation (%)	13.29	13.34	13.20	13.17	13.25	15.28	15.49	14.72
Lower Partial SD (%)	14.13	13.91	13.66	13.35	13.68	16.47	16.24	15.28
Avg. Gross Leverage	1.60	1.60	1.60	1.60	1.60	1.60	1.60	1.60
Prop. of Leverage (%)	1.91	3.11	3.08	3.06	2.93	0.71	0.85	1.41
Avg. Turnover (%)	0.43	0.40	0.42	0.41	0.42	0.09	0.10	0.11
Avg. Excess Return (%)	16.72	18.23	19.01	22.42	21.22	13.68	14.24	16.91
Cumulative Return (%)	34.88	38.74	40.83	50.14	46.79	26.74	27.99	34.86
Sharpe Ratio	1.26	1.37	1.44	1.70	1.60	0.90	0.92	1.15

	1 Factor VHAR (Log matrix)		3 Factors VHAR (Log matrix)		5 Factors VHAR (Log matrix)		7 Factors VHAR (Log matrix)	
	LASSO	aLASSO	LASSO	aLASSO	LASSO	aLASSO	LASSO	aLASSO
Standard Deviation (%)	13.20	13.37	12.81	12.86	12.57	12.83	12.63	12.75
Lower Partial SD (%)	13.29	13.64	12.60	12.54	12.54	12.75	12.52	12.62
Avg. Gross Leverage	1.60	1.60	1.60	1.60	1.60	1.60	1.60	1.60
Prop. of Leverage (%)	2.46	2.44	2.37	2.38	2.43	2.41	2.27	2.25
Avg. Turnover (%)	0.22	0.23	0.24	0.24	0.23	0.24	0.22	0.23
Avg. Excess Return (%)	16.07	19.89	19.72	21.04	20.56	18.93	20.74	19.19
Cumulative Return (%)	33.30	43.13	42.88	46.43	45.22	40.76	45.67	41.48
Sharpe Ratio	1.22	1.49	1.54	1.64	1.64	1.48	1.64	1.51

Portfolio Results

Statistics for Daily Portfolios - Restricted Minimum Variance (Long Only)

- No short-selling.
- The problem for an investor at time $t = t_0, \ldots, T-1$ is then given by

$$\boxed{\begin{aligned} \widehat{\boldsymbol{w}}_{t+1} = \arg\min_{\boldsymbol{w}_{t+1}} \quad & \boldsymbol{w}'_{t+1}\widehat{\boldsymbol{\Sigma}}_{t+1}\boldsymbol{w}_{t+1} \\ \text{subject to} \quad & \boldsymbol{w}'_{t+1}\mathbf{1} = 1, \\ & 0 \leq w_{it+1} \leq 0.20. \end{aligned}}$$

Portfolio Results

Statistics for Daily Portfolios - Restricted Minimum Variance (Long Only)

	RW	Block 1F	Block 3F	Block 5F	Block 7F	EWMA	BEKK-NL	DCC - NL
Standard Deviation (%)	17.10	17.06	16.96	16.85	16.88	17.74	17.92	17.78
Lower Partial SD (%)	17.56	17.83	17.63	17.49	17.58	18.94	19.16	19.13
Avg. Gross Leverage	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Prop. of Leverage (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg. Turnover (%)	0.17	0.16	0.16	0.16	0.16	0.03	0.03	0.04
Avg. Excess Return (%)	14.29	15.86	16.18	14.98	15.06	20.22	15.85	16.28
Cumulative Return (%)	27.49	31.30	32.15	29.25	29.44	42.18	30.91	32.04
Sharpe Ratio	0.84	0.93	0.95	0.89	0.89	1.14	0.88	0.92

	1 Factor VHAR (Log matrix)		3 Factors VHAR (Log matrix)		5 Factors VHAR (Log matrix)		7 Factors VHAR (Log matrix)	
	LASSO	aLASSO	LASSO	aLASSO	LASSO	aLASSO	LASSO	aLASSO
Standard Deviation (%)	16.96	16.98	16.55	16.59	16.34	16.47	16.31	16.44
Lower Partial Standard Deviation (%)	17.51	17.64	17.29	17.27	16.88	17.10	16.89	17.03
Prop. of Leverage (%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Avg. Turnover (%)	0.08	0.08	0.07	0.08	0.07	0.08	0.07	0.07
Avg. Excess Return (%)	17.60	17.57	17.62	18.04	18.02	18.17	17.13	17.04
Cumulative Return (%)	35.71	35.63	35.95	37.01	37.06	37.38	34.79	34.50
Sharpe Ratio	1.04	1.03	1.06	1.09	1.10	1.10	1.05	1.04