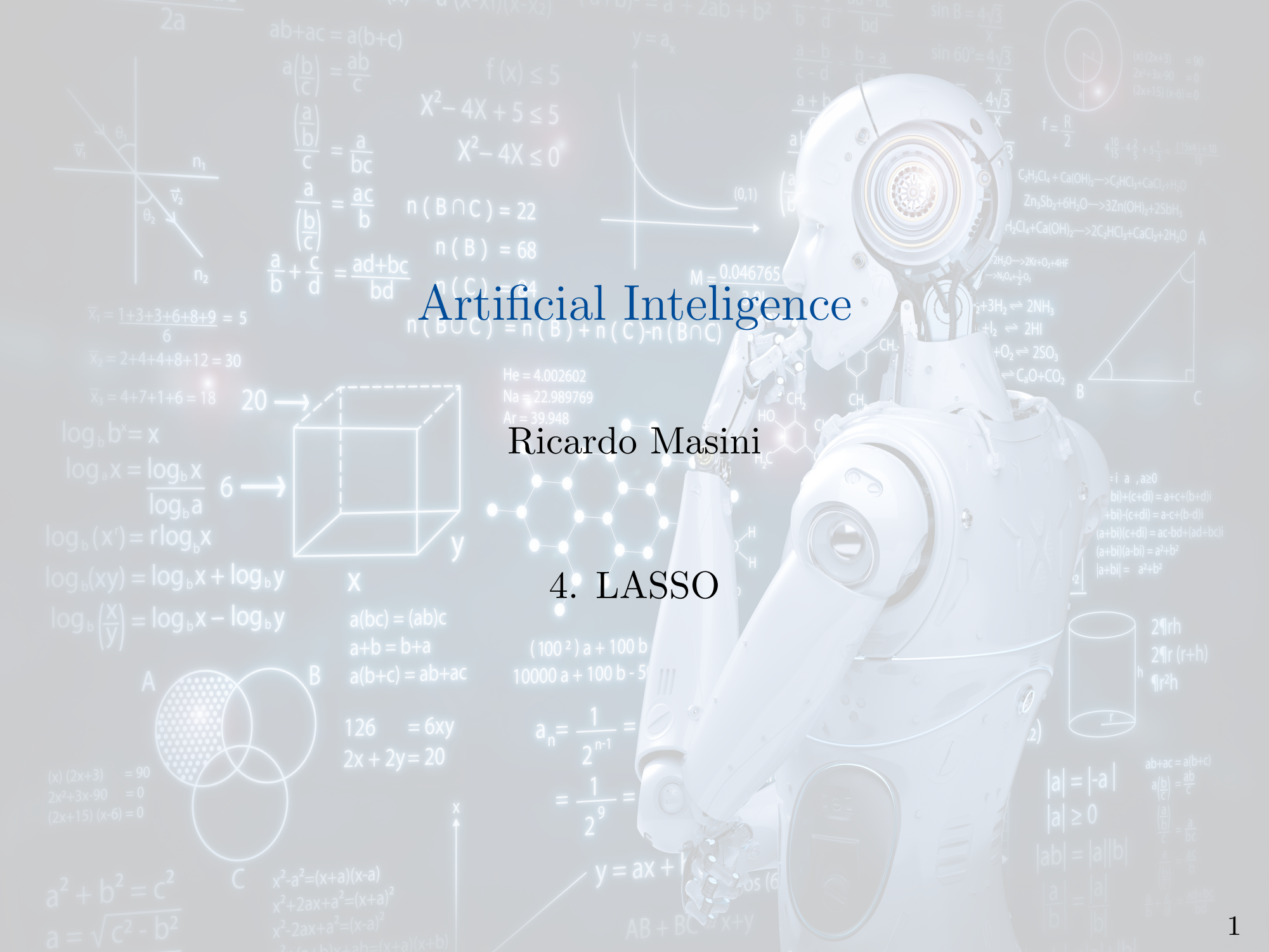


Artificial Inteligence

Ricardo Masini

4. LASSO

Agenda for Today

1. The LASSO
2. LASSO extensions: adaptive LASSO and Elastic Net
3. Solution to the LASSO Optimization problem.
4. Application: forecasting high-frequency returns
5. References:
   - Hastie, Tibshirani and Friedman (2009), Sections 3.4.2, 3.4.4 and 3.8

Shrinkage in Linear Models

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

p: número de regressores (preditores)
n: tamanho da amostra.

Framework: Linear Regression Model

Single-equation predictive linear models

$$Y_i = \beta_0 + \boldsymbol{\beta}' \boldsymbol{X}_i + U_i$$

- $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{pi})'$ is a vector of p regressors;
- U_i is a zero-mean innovation;
- $p \equiv p(n)$ and n is the sample size.

- Without loss of generality: $\beta_0 = 0$. Demeaned data.

Framework: Linear Regression Model

Single-equation predictive linear models

$$Y_i = \beta_0 + \boldsymbol{\beta}' \boldsymbol{X}_i + U_i$$

- $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{pi})'$ is a vector of p regressors;
- U_i is a zero-mean innovation;
- $p \equiv p(n)$ and n is the sample size.

- Without loss of generality: $\beta_0 = 0$. Demeaned data.
- All explanatory variables are in the same scale.

Antes de utilizar o LASSO, precisamos usar um operador de normalização

Framework: Linear Regression Model

Single-equation predictive linear models

$$Y_i = \beta_0 + \boldsymbol{\beta}' \boldsymbol{X}_i + U_i$$

- $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{pi})'$ is a vector of p regressors;
- U_i is a zero-mean innovation;
- $p \equiv p(n)$ and n is the sample size.

- Without loss of generality: $\beta_0 = 0$. Demeaned data.
- All explanatory variables are in the same scale.
- How should we estimate the parameters? Remember that p could be much larger than n.

Framework: Linear Regression Model

Single-equation predictive linear models

$$Y_i = \beta_0 + \boldsymbol{\beta}' \boldsymbol{X}_i + U_i$$

- $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{pi})'$ is a vector of p regressors;
- U_i is a zero-mean innovation;
- $p \equiv p(n)$ and n is the sample size.

- Without loss of generality: $\beta_0 = 0$. Demeaned data.
- All explanatory variables are in the same scale.
- How should we estimate the parameters? Remember that p could be much larger than n.
- What are the statistical properties of the estimated model?

Penalized Least Squares

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

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \sum_{j=1}^{p} p_{\lambda}(|\beta_j|; \boldsymbol{\alpha}, \text{data}) \right]$$

where

Penalized Least Squares

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

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data}) \right]$$

where

  - $p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data})$ is a non-negative penalty function indexed by the **regularization parameter** λ and could depend as well on the data and on additional hyper-parameters ($\boldsymbol{\alpha}$). E.g., $p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data}) = \lambda|\beta_j|^2$, $p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data}) = \lambda|\beta_j|$, or $p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data}) = \alpha\lambda|\beta_j| + (1-\alpha)\lambda|\beta_j|^2$.

Penalized Least Squares

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

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data}) \right]$$

  where
  - $p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data})$ is a non-negative penalty function indexed by the **regularization parameter** λ and could depend as well on the data and on additional hyper-parameters ($\boldsymbol{\alpha}$). E.g., $p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data}) = \lambda|\beta_j|^2$, $p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data}) = \lambda|\beta_j|$, or $p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data}) = \alpha\lambda|\beta_j| + (1-\alpha)\lambda|\beta_j|^2$.
- λ controls the "number of parameters" in the model.

Penalized Least Squares

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

$$\widehat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}\in\mathcal{B}} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data}) \right]$$

  where
  - $p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data})$ is a non-negative penalty function indexed by the **regularization parameter** λ and could depend as well on the data and on additional hyper-parameters ($\boldsymbol{\alpha}$). E.g., $p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data}) = \lambda|\beta_j|^2$, $p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data}) = \lambda|\beta_j|$, or $p_\lambda(|\beta_j|; \boldsymbol{\alpha}, \text{data}) = \alpha\lambda|\beta_j| + (1-\alpha)\lambda|\beta_j|^2$.
- 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.

Se o parâmetro de regularização for muito grande, a função penalidade é tão grande quanto, fazendo com que nenhuma variável seja "boa o suficiente" para entrar no modelo.

Penalized Least Squares

The bet on sparsity

- 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).

Supondo que temos uma quantidade candidatos a regressores (preditores) muito maior do que o tamanho da amostra, i.e. p >> n, apostar em esparsidade é apostar que um modelo, por exemplo o LASSO, vai selecionar uma quantidade muito pequena de regressores (preditores).

Penalized Least Squares

The bet on sparsity

- 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).
  - $\boldsymbol{X}_i = [\boldsymbol{X}'_{i,S}, \boldsymbol{X}'_{i,S^c}]'$, $\boldsymbol{X}_{i,S} \in \mathbb{R}^s$ is the vector of **relevant** variables and $\boldsymbol{X}_{i,S^c} \in \mathbb{R}^{p-s}$ is the vector of **irrelevant** ones. $\boldsymbol{\beta} = [\boldsymbol{\beta}'_S, \boldsymbol{\beta}'_{S^c}]'$. Isso significa que temos apenas S preditores relevantes.

Penalized Least Squares

The bet on sparsity

- 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).
  - $\boldsymbol{X}_i = [\boldsymbol{X}'_{i,S}, \boldsymbol{X}'_{i,S^c}]'$, $\boldsymbol{X}_{i,S} \in \mathbb{R}^s$ is the vector of **relevant** variables and $\boldsymbol{X}_{i,S^c} \in \mathbb{R}^{p-s}$ is the vector of **irrelevant** ones. $\boldsymbol{\beta} = [\boldsymbol{\beta}'_S, \boldsymbol{\beta}'_{S^c}]'$.

- 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. $s << p$

Penalized Least Squares

The bet on sparsity

- 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).
  - $\boldsymbol{X}_i = [\boldsymbol{X}'_{i,S}, \boldsymbol{X}'_{i,S^c}]'$, $\boldsymbol{X}_{i,S} \in \mathbb{R}^s$ is the vector of **relevant** variables and $\boldsymbol{X}_{i,S^c} \in \mathbb{R}^{p-s}$ is the vector of **irrelevant** ones. $\boldsymbol{\beta} = [\boldsymbol{\beta}'_S, \boldsymbol{\beta}'_{S^c}]'$.

- 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 LASSO

Tibshirani (JRRS B, 1996)

- Least Absolute Shrinkage and Selection Operator (LASSO):

$$\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

Isso é, dentro da classe de modelos de mínimos quadrados penalizados, o LASSO é um que utiliza a função penalidade mais simples, p_lamba = lambda modulo(beta_j).

The LASSO

Tibshirani (JRRS B, 1996)

- Least Absolute Shrinkage and Selection Operator (LASSO):

$$\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

- "Shrinks" to zero parameters associated with redundant predictors. "Encolhe" para zero os parâmetros associados à preditores redundantes.

The LASSO

Tibshirani (JRRS B, 1996)

- Least Absolute Shrinkage and Selection Operator (LASSO):

$$\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

- "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):

$$\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

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

The LASSO

Tibshirani (JRRS B, 1996)

- Least Absolute Shrinkage and Selection Operator (LASSO):

$$\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]$$

- "Shrinks" to zero parameters associated with redundant predictors.
- The regularization path can be efficiently estimated.
- Can handle (many) more variables than observations $(p >> n)$. muito mais variáveis, não necessáriamente variáveis relevantes.
- Under some conditions can select the correct subset of relevant variables. S*

The LASSO

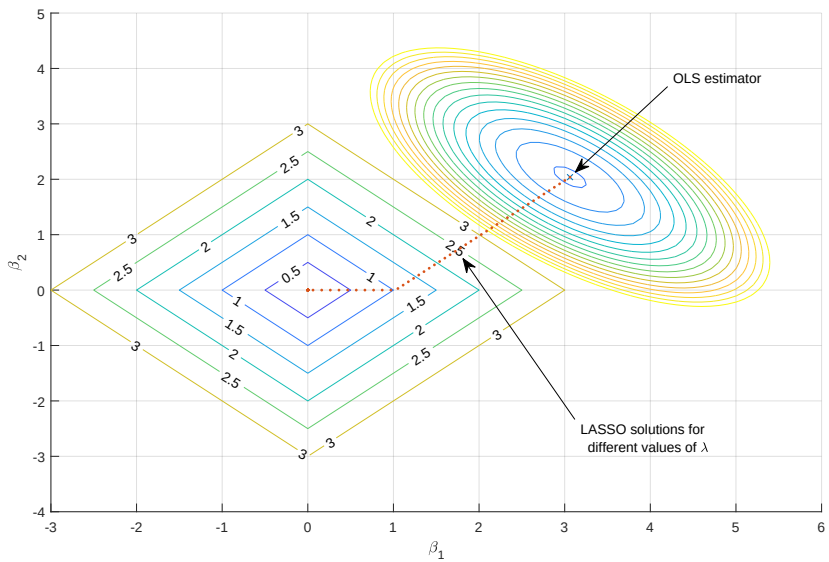
LASSO as a Constrained Optimization Program

$$\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \sum_{i=1}^{n}(Y_i - \boldsymbol{\beta}'\boldsymbol{X}_i)^2$$

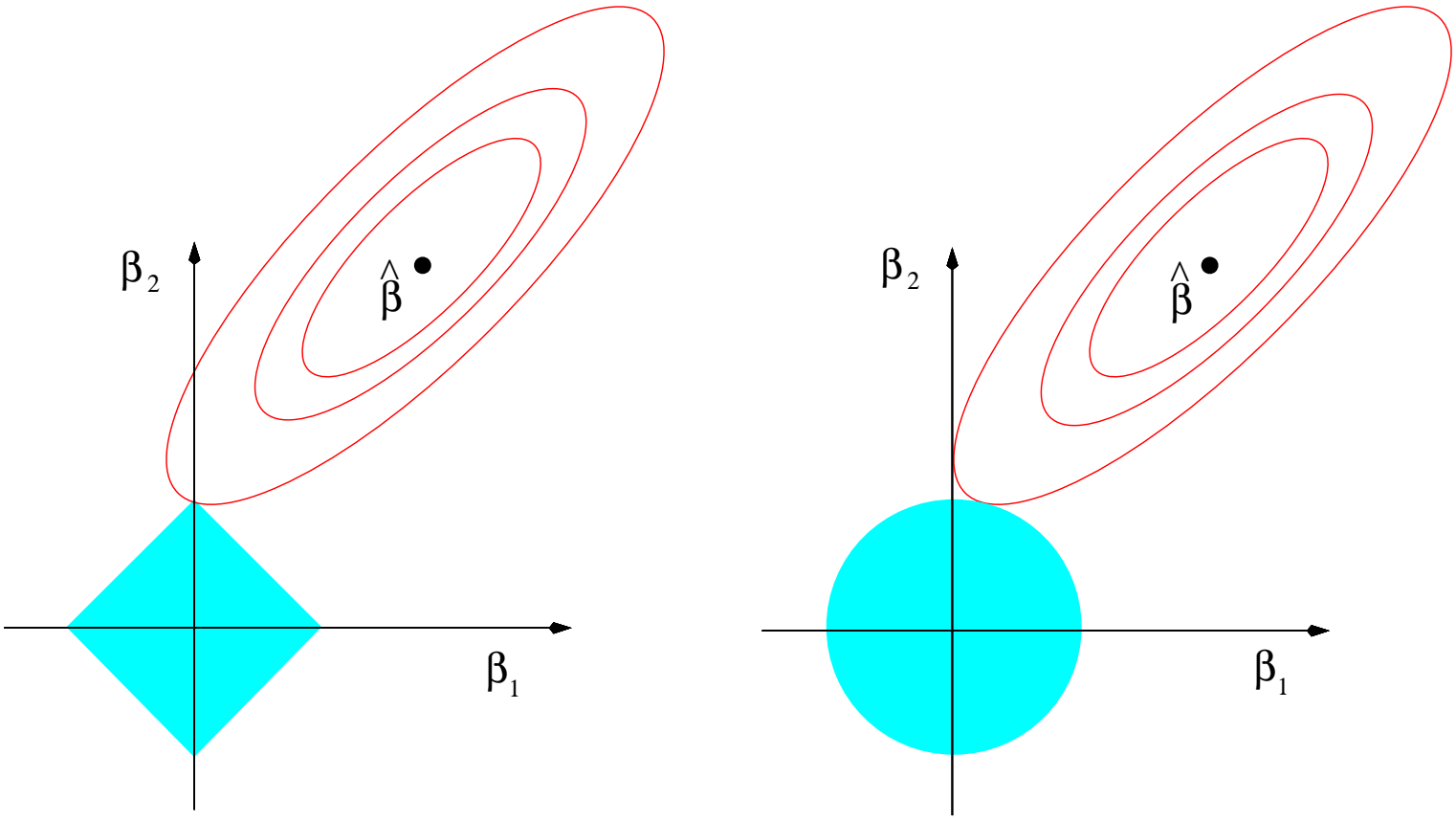
$$\text{subject to}$$

$$\sum_{j=1}^{p}|\beta_j| < c(\lambda)$$

- The figure represents the sum-of-squared errors level curves as well as the LASSO restrictions.
- The cross indicates the OLS (unrestricted) solution.
- Red dots represent the LASSO solutions as a function of λ.

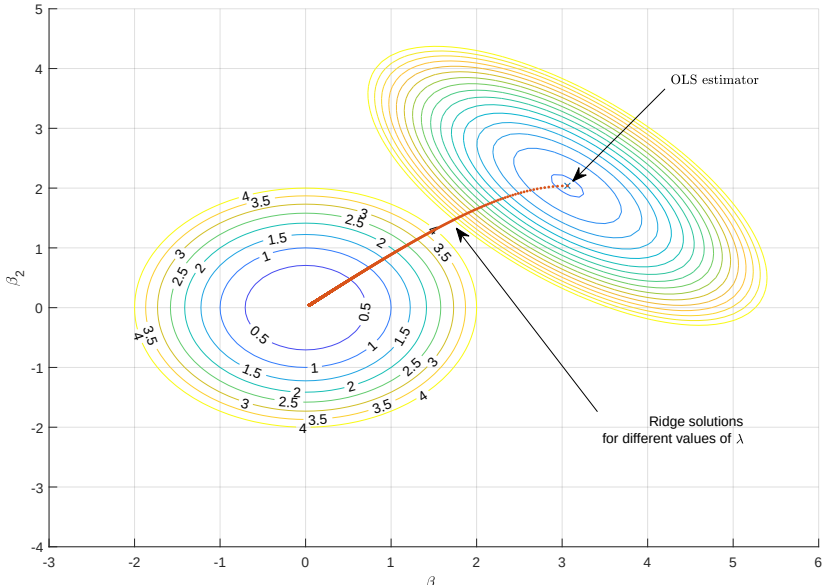


LASSO versus Ridge

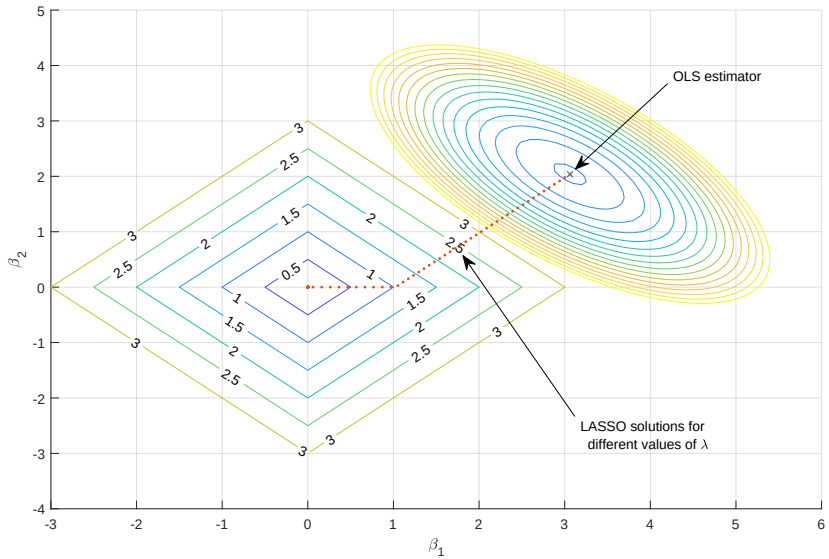


LASSO versus Ridge

Ridge



LASSO



LASSO and Model Selection

Consistency

Estimation Consistency

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

Model Selection Consistency

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

Sign Consistency

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

where

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

LASSO and Model Selection

The sign Function

The sign function is defined as

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

Sign Consistency

Definitions

Strong Sign Consistency

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

$$\lim_{n \to \infty} \mathsf{P}\left[\widehat{\boldsymbol{\beta}}(\lambda_n) \overset{s}{=} \boldsymbol{\beta}^0\right] = 1$$

General Sign Consistency

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

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

- 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. $\widehat{\boldsymbol{\Sigma}}$ is the empirical covariance matrix of regressors. For example: $\widehat{\boldsymbol{\Sigma}}_{S^c S} = (1/n)\boldsymbol{X}'_{S^c}\boldsymbol{X}_S$.
- The Irrepresentable Condition is a key condition for model selection consistency of the LASSO!

The Adaptive LASSO - Zou (JASA, 2006)

- The Adaptive LASSO (adaLASSO) estimator:

$$\widehat{\boldsymbol{\beta}}_{adaLASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right]$$

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

The Adaptive LASSO - Zou (JASA, 2006)

- The Adaptive LASSO (adaLASSO) estimator:

$$\widehat{\boldsymbol{\beta}}_{adaLASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right]$$

  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:

$$\widehat{\boldsymbol{\beta}}_{adaLASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right]$$

  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:

$$\widehat{\boldsymbol{\beta}}_{adaLASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right]$$

  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 Elastic Net - Zou and Hastie (JRRS B, 2005)

- The Elastic Net estimator is defined as

$$\widehat{\boldsymbol{\beta}}_{EN}(\lambda) = \underset{\boldsymbol{\beta}\in\beta}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \alpha\lambda \sum_{j=1}^{p} \beta_i^2 + (1-\alpha)\lambda \sum_{j=1}^{p} |\beta_i| \right]$$

The Elastic Net - Zou and Hastie (JRRS B, 2005)

- The Elastic Net estimator is defined as

$$\widehat{\boldsymbol{\beta}}_{EN}(\lambda) = \underset{\boldsymbol{\beta}\in\beta}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \alpha\lambda \sum_{j=1}^{p} \beta_i^2 + (1-\alpha)\lambda \sum_{j=1}^{p} |\beta_i| \right]$$

- Motivation: correlated regressors.

The Elastic Net - Zou and Hastie (JRRS B, 2005)

- The Elastic Net estimator is defined as

$$\widehat{\boldsymbol{\beta}}_{EN}(\lambda) = \underset{\boldsymbol{\beta}\in\beta}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \alpha\lambda \sum_{j=1}^{p} \beta_i^2 + (1-\alpha)\lambda \sum_{j=1}^{p} |\beta_i| \right]$$

- Motivation: correlated regressors.
- The ℓ_1-part of the penalty generates a sparse model.

The Elastic Net - Zou and Hastie (JRRS B, 2005)

- The Elastic Net estimator is defined as

$$\widehat{\boldsymbol{\beta}}_{EN}(\lambda) = \underset{\boldsymbol{\beta}\in\beta}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \alpha\lambda \sum_{j=1}^{p} \beta_i^2 + (1-\alpha)\lambda \sum_{j=1}^{p} |\beta_i| \right]$$

- Motivation: correlated regressors.
- The ℓ_1-part of the penalty generates a sparse model.
- The quadratic part of the penalty:
  - Removes the limitation on the number of selected variables;
  - Encourages grouping effect;
  - Stabilizes the ℓ_1-regularization path. $0 \leq \alpha \leq 1$

The Elastic Net - Zou and Hastie (JRRS B, 2005)

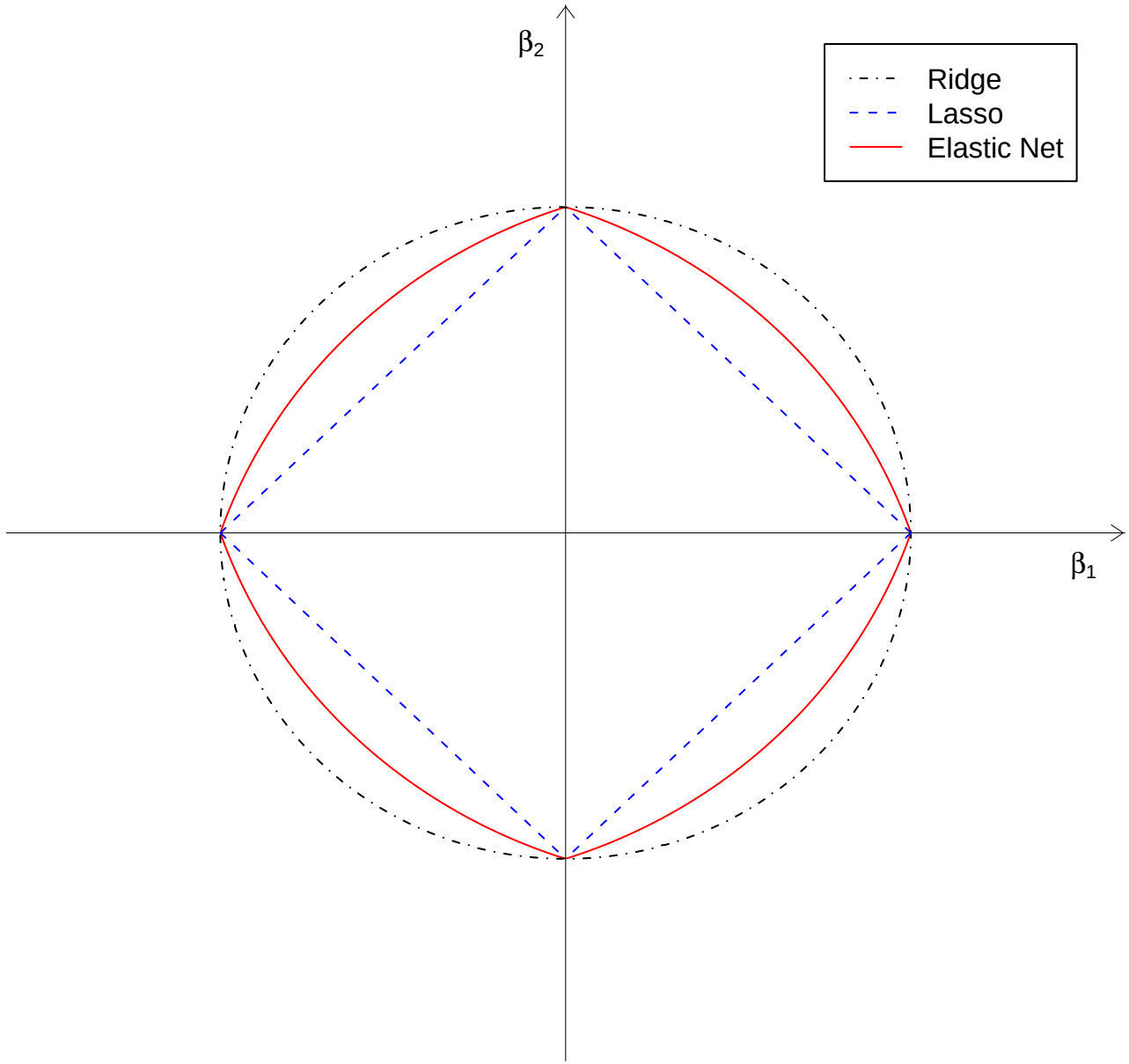
- The Elastic Net estimator is defined as

$$\widehat{\boldsymbol{\beta}}_{EN}(\lambda) = \underset{\boldsymbol{\beta}\in\beta}{\arg\min} \left[\sum_{i=1}^{n} (Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i)^2 + \alpha\lambda \sum_{j=1}^{p} \beta_i^2 + (1-\alpha)\lambda \sum_{j=1}^{p} |\beta_i| \right]$$

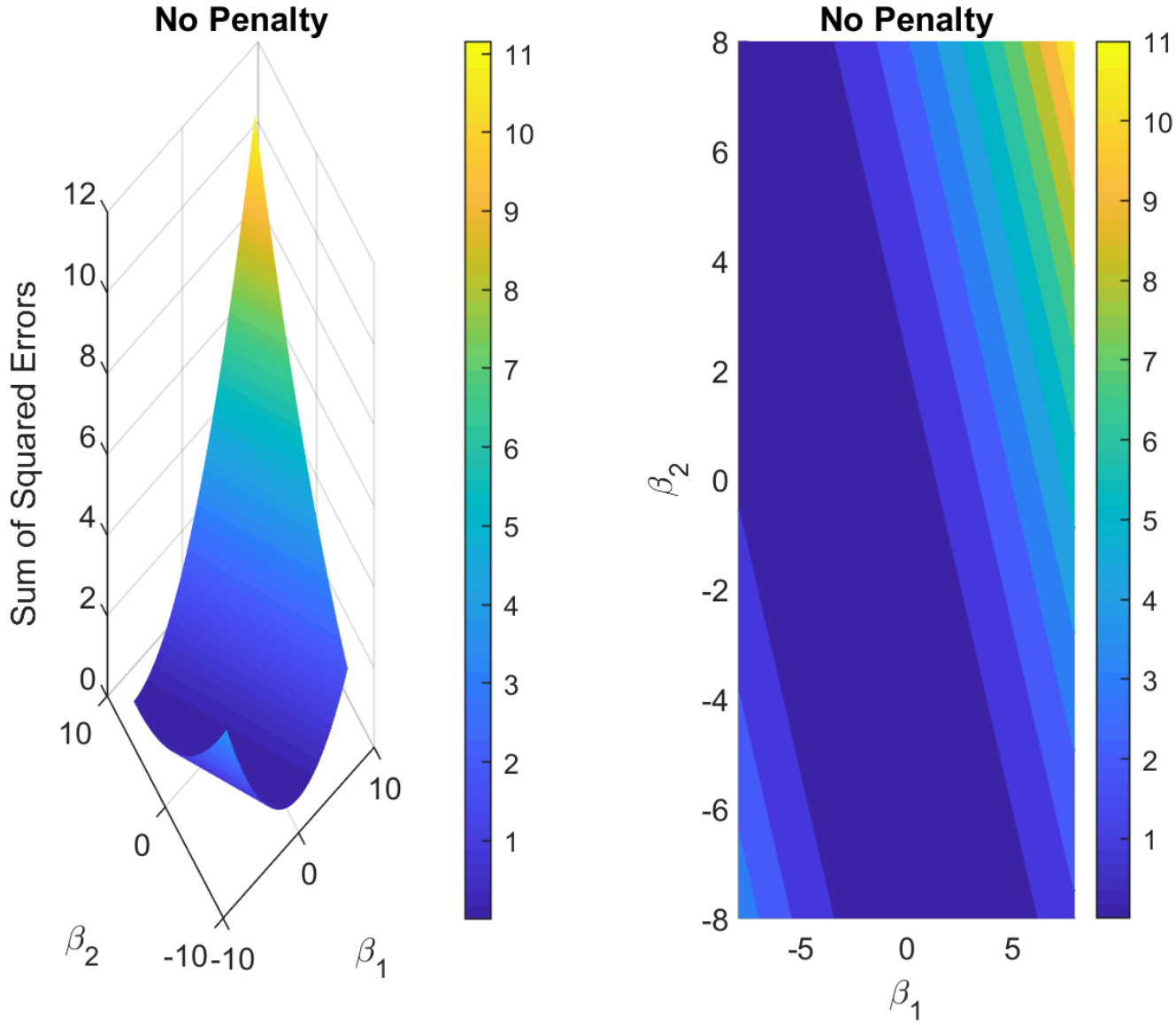
- Motivation: correlated regressors.
- The ℓ_1-part of the penalty generates a sparse model.
- The quadratic part of the penalty:
  - Removes the limitation on the number of selected variables;
  - Encourages grouping effect;
  - Stabilizes the ℓ_1-regularization path. $0 \leq \alpha \leq 1$
- Adaptive EL-Net was proposed by Zou and Zhang (AoS, 2009).

The Elastic Net Estimator

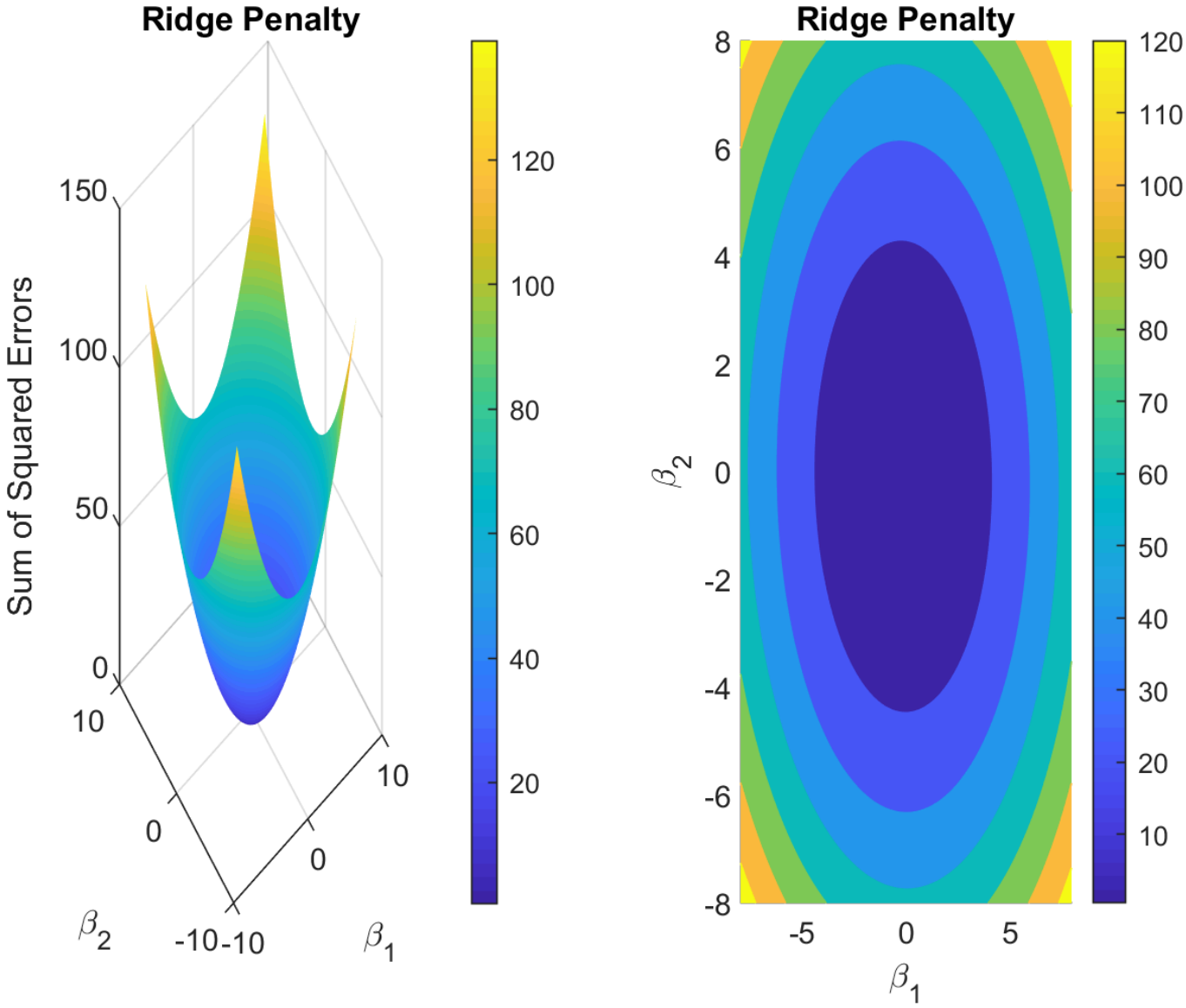
The Geometry of the Elastic Net



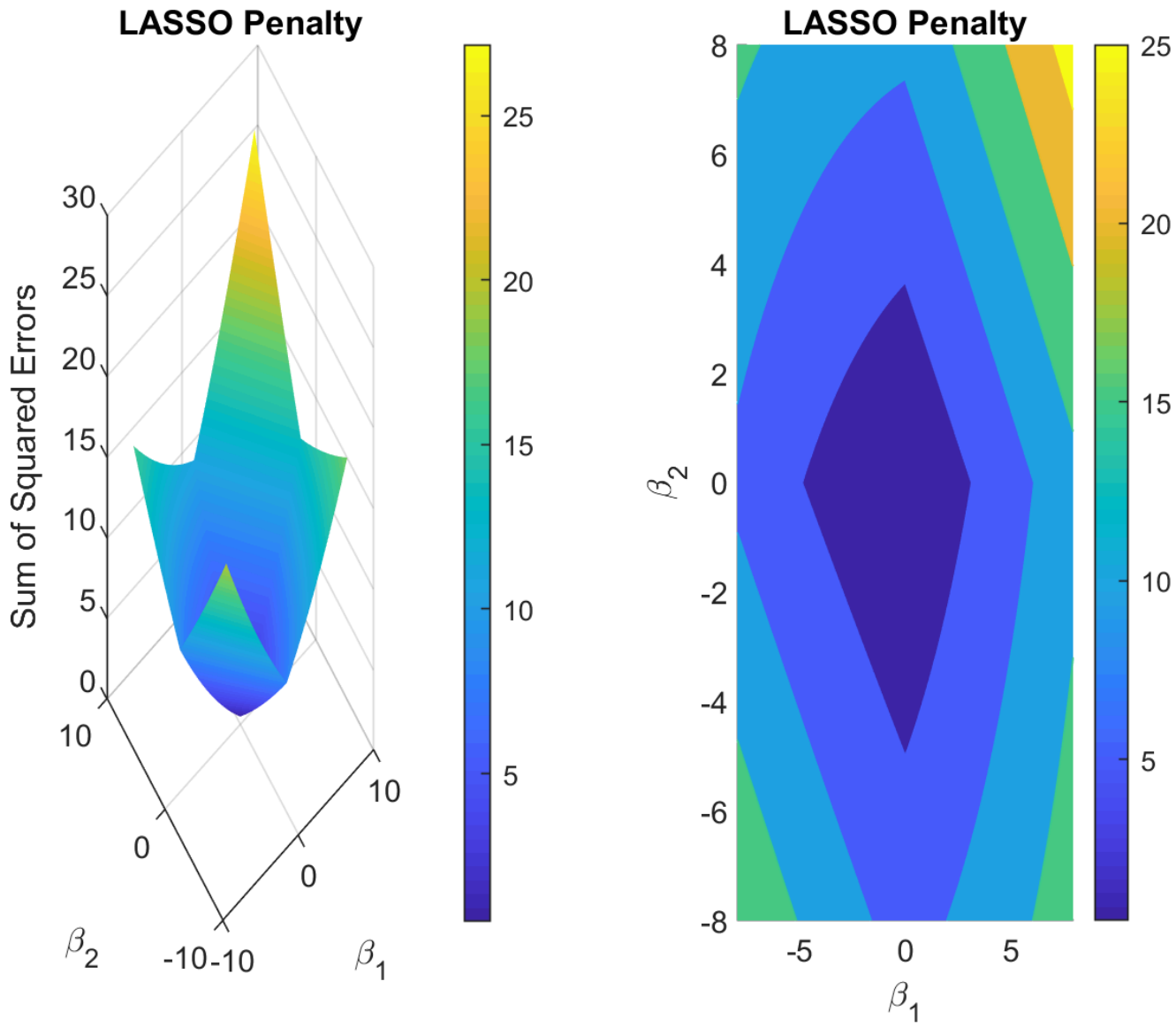
Why Does Shrinkage Work When $p > n$



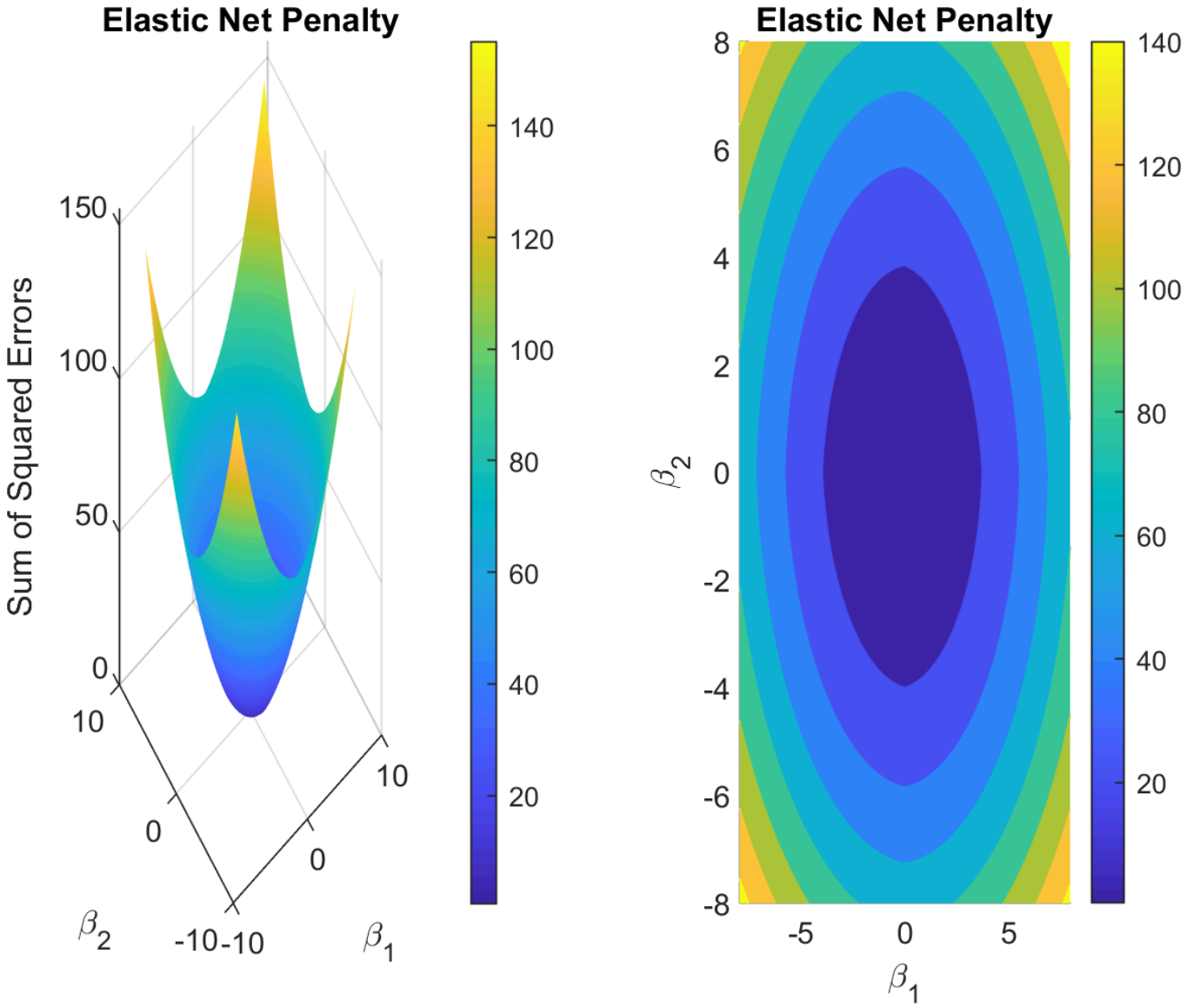
Why Does Shrinkage Work When $p > n$



Why Does Shrinkage Work When $p > n$



Why Does Shrinkage Work When $p > n$



Solution for the LASSO

Solution for the LASSO Program

- The LASSO problem is a ***convex program***: quadratic program with a convex constraint.

Solution for the LASSO Program

- The LASSO problem is a ***convex program***: quadratic program with a convex constraint.

- However, it is not ***strictly convex*** when $\boldsymbol{Z}'\boldsymbol{Z}$ is singular. Therefore,

$$\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) \text{ is, in general, not unique.}$$
$$\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) \text{ is unique.}$$

Solution for the LASSO Program

- The LASSO problem is a ***convex program***: quadratic program with a convex constraint.

- However, it is not ***strictly convex*** when $\boldsymbol{Z}'\boldsymbol{Z}$ is singular. Therefore,

$$\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) \text{ is, in general, not unique.}$$
$$\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) \text{ is unique.}$$

- Luckily, Tibshirani (EJS, 2013) proves that:

  If the entries of $\boldsymbol{Z}$ are drawn from a continuous probability distribution, than the LASSO solution is unique with probability one.

Solution for the LASSO Program

- The LASSO problem is a ***convex program***: quadratic program with a convex constraint.

- However, it is not ***strictly convex*** when $\boldsymbol{Z}'\boldsymbol{Z}$ is singular. Therefore,

  $\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)$ is, in general, not unique.

  $\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)$ is unique.

- Luckily, Tibshirani (EJS, 2013) proves that:

  If the entries of $\boldsymbol{Z}$ are drawn from a continuous probability distribution, than the LASSO solution is unique with probability one.

- The number of nonzero parameter estimates will be at most $\mathsf{min}(n, p)$.

Solution of the LASSO Program

- Any LASSO solution $\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)$ must satisfy

$$\boxed{\frac{1}{n}\boldsymbol{Z}'[\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)] = \lambda \boldsymbol{sg},}$$

where $\boldsymbol{sg} = (sg_1, \ldots, sg_p)'$ and, for $j = 1, \ldots, p$:

$$sg_j \in \begin{cases} \{+1\} & \text{if } \widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) > 0 \\ \{\text{-}1\} & \text{if } \widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) < 0 \\ [\text{-}1,1] & \text{if } \widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = 0 \end{cases}$$

Solution of the LASSO Program

- Any LASSO solution $\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)$ must satisfy

$$\boxed{\frac{1}{n}\boldsymbol{Z}'[\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)] = \lambda \boldsymbol{sg},}$$

where $\boldsymbol{sg} = (sg_1, \ldots, sg_p)'$ and, for $j = 1, \ldots, p$:

$$sg_j \in \begin{cases} \{+1\} & \text{if } \widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) > 0 \\ \{\text{-}1\} & \text{if } \widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) < 0 \\ [\text{-}1,1] & \text{if } \widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = 0 \end{cases}$$

- Why is this the case? Subgradient optimality.

Solution of the LASSO Program

- Any LASSO solution $\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)$ must satisfy

$$\boxed{\frac{1}{n}\boldsymbol{Z}'[\boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)] = \lambda \boldsymbol{sg},}$$

where $\boldsymbol{sg} = (sg_1, \ldots, sg_p)'$ and, for $j = 1, \ldots, p$:

$$sg_j \in \begin{cases} \{+1\} & \text{if } \widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) > 0 \\ \{\text{-}1\} & \text{if } \widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) < 0 \\ [\text{-}1,1] & \text{if } \widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) = 0 \end{cases}$$

- Why is this the case? Subgradient optimality.
- $\boldsymbol{sg} \in \partial \left\| \widehat{\boldsymbol{\beta}}_{LASSO}(\lambda) \right\|_1$, a subgradient of the ℓ_1 norm of $\boldsymbol{\beta}$ evaluated at $\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)$.

Solution of the LASSO Program

Subgradient and Subderivative

- The subdifferential $\partial f(x_0)$ is the set $[\mathsf{d}_- f(x_0), \mathsf{d}_+ f(x_0)]$, where $\mathsf{d}_- f(x_0)$ and $\mathsf{d}_- f(x_0)$ are one-sided limits (left- and right-derivatives of $f(x)$ at x_0):

$$\boxed{\begin{aligned} \mathsf{d}_- f(x_0) &= \lim_{x \to x_0^-} \frac{f(x) - f(x_0)}{x - x_0} \\ \mathsf{d}_+ f(x_0) &= \lim_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0}. \end{aligned}}$$

Solution of the LASSO Program

Subgradient and Subderivative

- The subdifferential $\partial f(x_0)$ is the set $[\mathsf{d}_- f(x_0), \mathsf{d}_+ f(x_0)]$, where $\mathsf{d}_- f(x_0)$ and $\mathsf{d}_- f(x_0)$ are one-sided limits (left- and right-derivatives of $f(x)$ at x_0):

$$\boxed{\begin{aligned} \mathsf{d}_- f(x_0) &= \lim_{x \to x_0^-} \frac{f(x) - f(x_0)}{x - x_0} \\ \mathsf{d}_+ f(x_0) &= \lim_{x \to x_0^+} \frac{f(x) - f(x_0)}{x - x_0}. \end{aligned}}$$

- The subdifferential is a set-valued function: it consists of a single value if f is differentiable, an interval of values, or it is empty if $\mathsf{d}_- f(x) > \mathsf{d}_+ f(x)$.

Solution of the LASSO Program

Subgradient and Subderivative

- The essential results of optimization can be extended to semi-differentiable functions

Solution of the LASSO Program

Subgradient and Subderivative

- The essential results of optimization can be extended to semi-differentiable functions
- Result: If f is a semi-differentiable function and x_0 is a local minimum of f, then $0 \in \partial f(x_0)$.

Solution of the LASSO Program

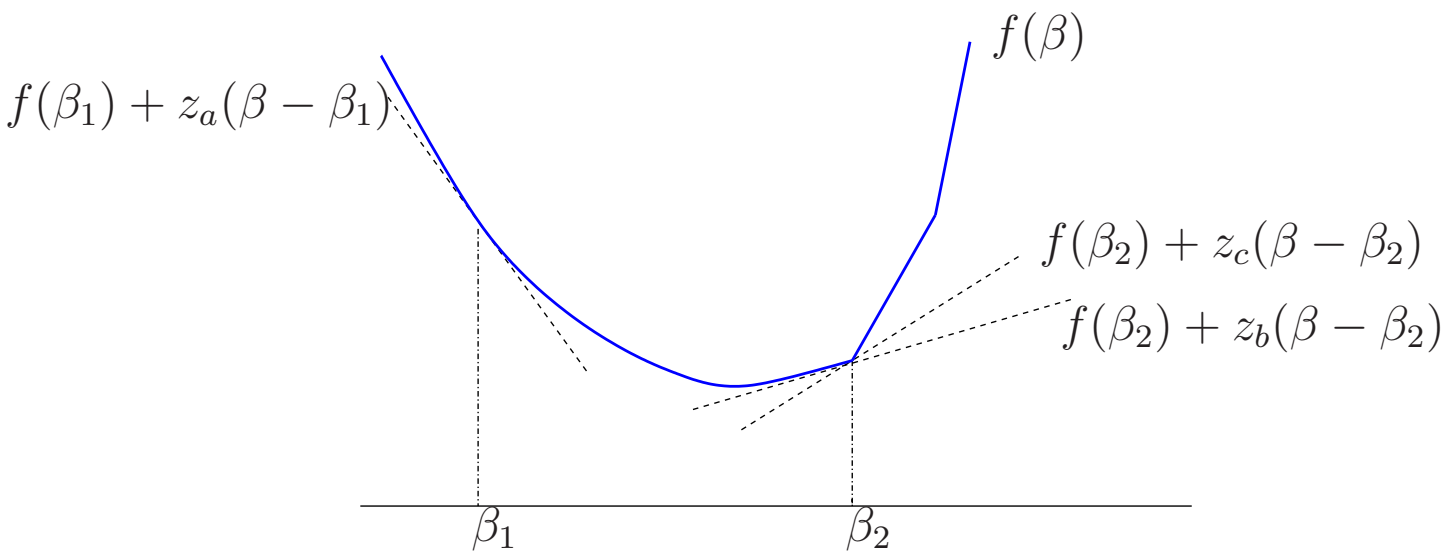
Subgradient and Subderivative

- The essential results of optimization can be extended to semi-differentiable functions
- Result: If f is a semi-differentiable function and x_0 is a local minimum of f, then $0 \in \partial f(x_0)$.
- As with regular calculus, the converse is not true in general.

Solution of the LASSO Program

Subgradient and Subderivative

- The essential results of optimization can be extended to semi-differentiable functions
- Result: If f is a semi-differentiable function and x_0 is a local minimum of f, then $0 \in \partial f(x_0)$.
- As with regular calculus, the converse is not true in general.



Solution of the LASSO Program

Subgradient and Subderivative

- For the absolute value function $f(x) = |x|$, the subderivative is given as

$$\partial f(x) = \begin{cases} -1 & \text{if } x < 0 \\ [-1, 1] & \text{if } x = 0 \\ 1 & \text{if } x > 0. \end{cases}$$

Solution of the LASSO Program

Karush-Kuhn-Tucker (KKT) Conditions

- $\widehat{\beta}_{LASSO}(\lambda)$ minimizes the LASSO objective function if and only if it satisfies the KKT (first-order) conditions:

$$\frac{1}{n}\frac{\partial \|\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}\|_2^2}{\partial \beta_j}\bigg|_{\beta_j=\widehat{\beta}_{LASSO,j}(\lambda)} = \lambda \mathsf{sign}[\widehat{\beta}_{LASSO,j}(\lambda)] \text{ if } \widehat{\beta}_{LASSO,j}(\lambda) \neq 0$$

$$\frac{1}{n}\left|\frac{\partial \|\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}\|_2^2}{\partial \beta_j}\right|_{\beta_j=\widehat{\beta}_{LASSO,j}(\lambda)} \leq \lambda \qquad \text{if } \widehat{\beta}_{LASSO,j}(\lambda) = 0.$$

Solution of the LASSO Program

Karush-Kuhn-Tucker (KKT) Conditions

- $\widehat{\beta}_{LASSO}(\lambda)$ minimizes the LASSO objective function if and only if it satisfies the KKT (first-order) conditions:

$$\frac{1}{n}\frac{\partial\|\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}\|_2^2}{\partial\beta_j}\bigg|_{\beta_j=\widehat{\beta}_{LASSO,j}(\lambda)} = \lambda\mathsf{sign}[\widehat{\beta}_{LASSO,j}(\lambda)] \text{ if } \widehat{\beta}_{LASSO,j}(\lambda)\neq 0$$

$$\frac{1}{n}\left|\frac{\partial\|\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\beta}\|_2^2}{\partial\beta_j}\right|_{\beta_j=\widehat{\beta}_{LASSO,j}(\lambda)} \leq \lambda \qquad \text{if } \widehat{\beta}_{LASSO,j}(\lambda)=0.$$

- Therefore,

$$\frac{1}{n}\boldsymbol{Z}_j'[\boldsymbol{Y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)] = \lambda\mathsf{sign}[\widehat{\beta}_{LASSO,j}(\lambda)] \text{ if } \widehat{\beta}_j(\lambda)\neq 0$$

$$\frac{1}{n}\left|\boldsymbol{Z}_j'[\boldsymbol{Y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)]\right| \leq \lambda \qquad \text{if } \widehat{\beta}_{LASSO,j}(\lambda)=0$$

Solution of the LASSO Program

Karush-Kuhn-Tucker (KKT) Conditions

- In other words, the correlation between a predictor X_{jt} and the residuals, $Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)$ must exceed a certain minimum threshold λ before it is included in the model.

Solution of the LASSO Program

Karush-Kuhn-Tucker (KKT) Conditions

- In other words, the correlation between a predictor X_{jt} and the residuals, $Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)$ must exceed a certain minimum threshold λ before it is included in the model.
- If this correlation is bellow λ, than $\widehat{\beta}_{LASSO,j}(\lambda) = 0$.

Solution of the LASSO Program

Karush-Kuhn-Tucker (KKT) Conditions

- In other words, the correlation between a predictor X_{jt} and the residuals, $Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)$ must exceed a certain minimum threshold λ before it is included in the model.
- If this correlation is bellow λ, than $\widehat{\beta}_{LASSO,j}(\lambda) = 0$.
- If $p \leq n$ and $\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{I}$, the estimator is the **soft threshold estimator**:

$$\widehat{\beta}_{LASSO,j}(\lambda) = \text{sign}(W_j)\left(|W_j| - \frac{\lambda}{2}\right)_+ := S_\lambda(W_j),$$

where $(x)_+ = \max(x, 0)$ and $W_j = \frac{\boldsymbol{Z}_j'\boldsymbol{Y}}{n}$.

Solution of the LASSO Program

Karush-Kuhn-Tucker (KKT) Conditions

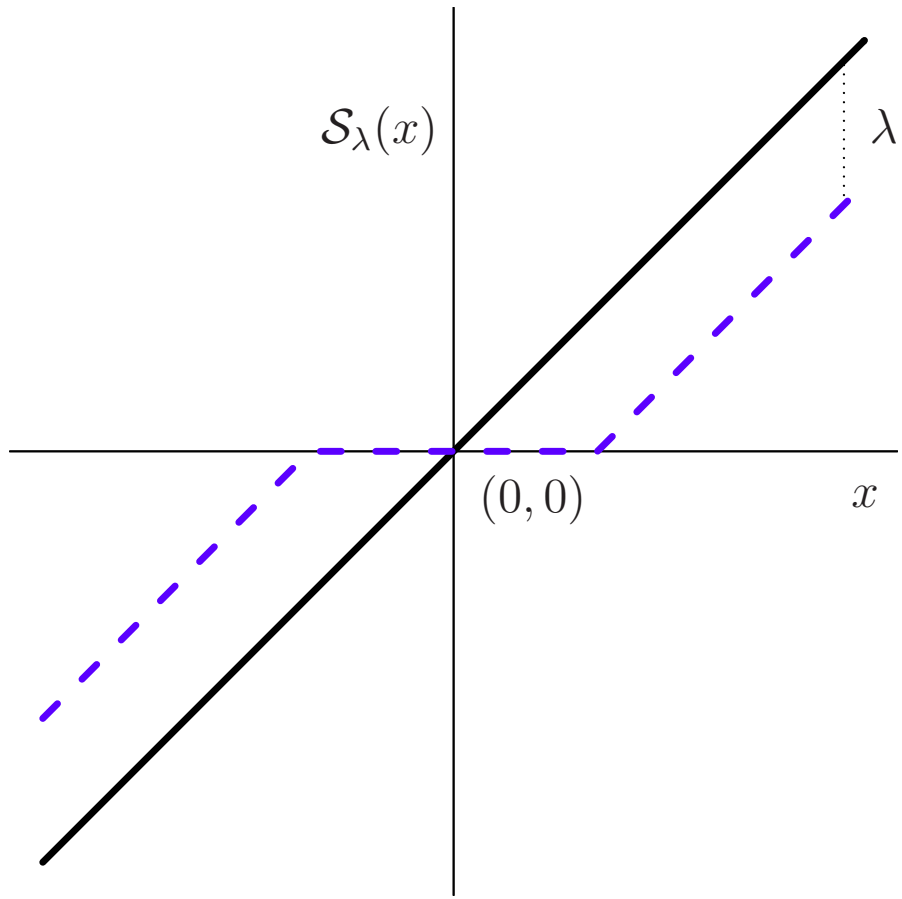
- In other words, the correlation between a predictor X_{jt} and the residuals, $Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)$ must exceed a certain minimum threshold λ before it is included in the model.
- If this correlation is bellow λ, than $\widehat{\beta}_{LASSO,j}(\lambda) = 0$.
- If $p \leq n$ and $\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z} = \boldsymbol{I}$, the estimator is the **soft threshold estimator**:

$$\widehat{\beta}_{LASSO,j}(\lambda) = \text{sign}(W_j)\left(|W_j| - \frac{\lambda}{2}\right)_+ := S_\lambda(W_j),$$

  where $(x)_+ = \max(x, 0)$ and $W_j = \frac{\boldsymbol{Z}_j'\boldsymbol{Y}}{n}$.
- Note that, in this case, $\frac{\boldsymbol{Z}_j\boldsymbol{Y}}{n}$ is the OLS estimator.

Solution of the LASSO Program

Orthonormal Case



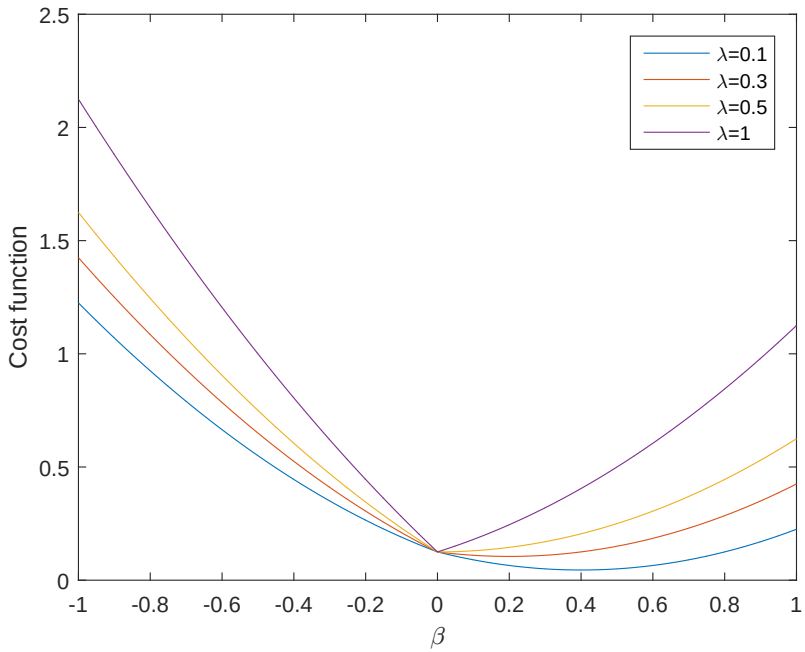
Solution of the LASSO Program

Orthonormal Case

- Hence,

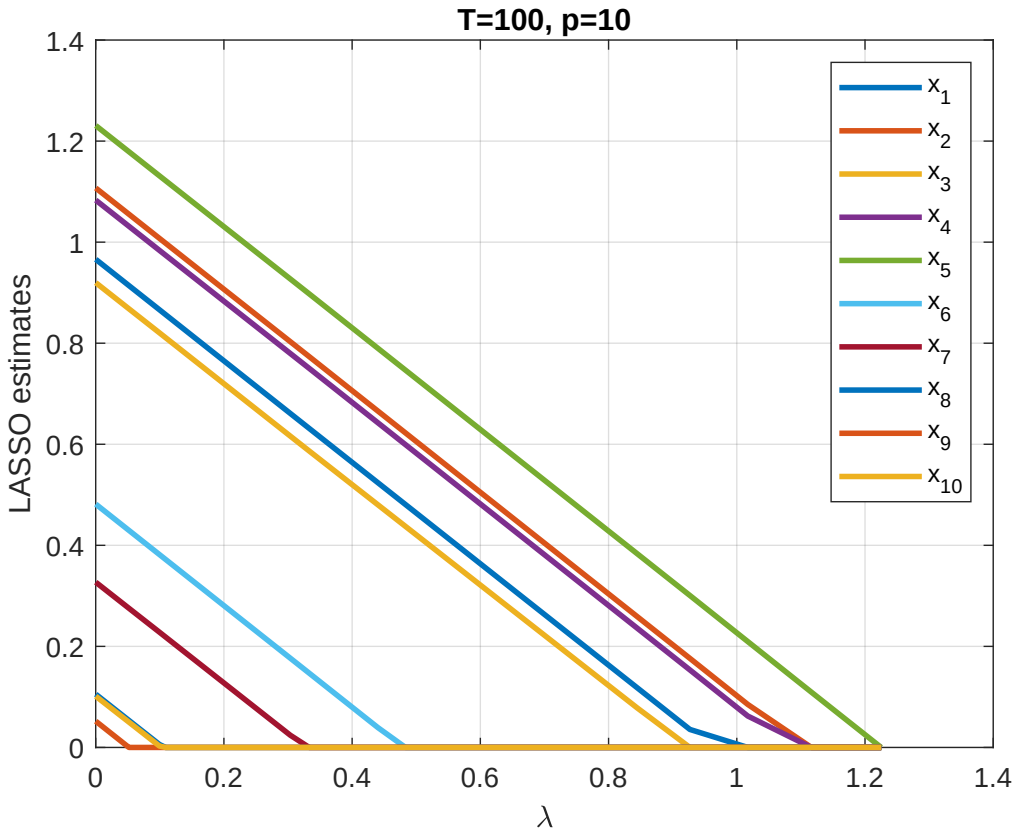
$$\widehat{\beta}_{LASSO}(\lambda) = \begin{cases} \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Y} - \lambda/2 & \text{if } \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Y} > \lambda/2 \\ 0 & \text{if } \frac{1}{T}|\boldsymbol{X}'\boldsymbol{Y}| \leq \lambda/2 \\ \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Y} + \lambda/2 & \text{if } \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Y} < -\lambda/2 \end{cases}$$

Example: $\beta = 0.5$



Solution of the LASSO Program

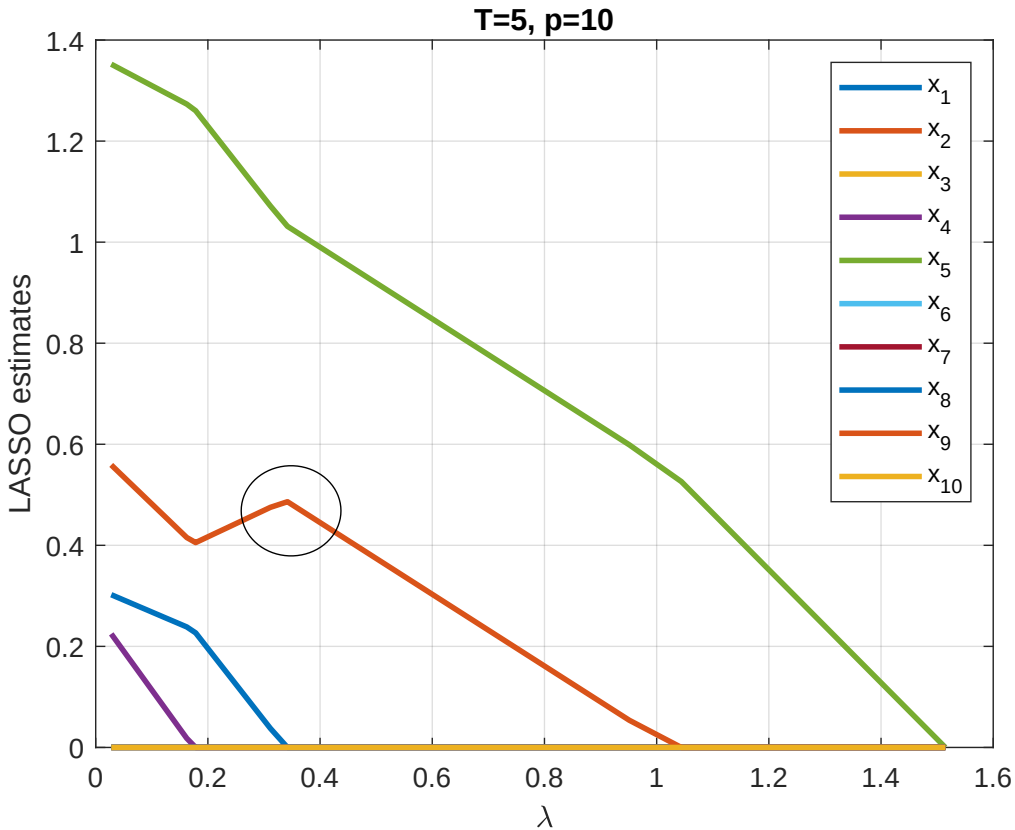
The LASSO Path: Estimates as a function of λ



- Simulated data from a linear regression model with $p = 10$ covariates. Number of observations: $T = 100$.
- Both $\boldsymbol{Z}$ and $\boldsymbol{U}$ come jointly from a $\mathsf{N}(\boldsymbol{0}, \boldsymbol{I})$ distribution.
- The figure shows the LASSO estimates as a function of λ.
- FACT 1: The LASSO path is piecewise linear.
- FACT 2: if $\lambda \geq \frac{1}{n}\|\boldsymbol{Z}'y\|_\infty$ all LASSO estimates will be zero. Therefore, $\lambda_{max} = \frac{1}{n}\|\boldsymbol{Z}'y\|_\infty$.

Solution of the LASSO Program

The LASSO Path: Estimates as a function of λ



- Simulated data from a linear regression model with $p = 10$ covariates. Number of observations: $T = 5$.
- Both $\boldsymbol{Z}$ and $\boldsymbol{U}$ come jointly from a $\mathsf{N}(\mathbf{0}, \boldsymbol{I})$ distribution.
- The figure shows the LASSO estimates as a function of λ.
- FACT 1: The LASSO path is piecewise linear.
- FACT 2: if $\lambda \geq \frac{1}{n}\|\boldsymbol{Z}'y\|_\infty$ all LASSO estimates will be zero. Therefore, $\lambda_{max} = \frac{1}{n}\|\boldsymbol{Z}'y\|_\infty$.
- Note that LASSO estimates can increase as a function of λ! Why?

Solution of the LASSO Program

The Active Set

- First, assume that $\boldsymbol{Z}$ is such that the LASSO solution is unique.

Solution of the LASSO Program

The Active Set

- First, assume that $\boldsymbol{Z}$ is such that the LASSO solution is unique.
- Let $A = \mathsf{supp}[\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)]$ be the LASSO active set (the nonzero coefficient estimates).

Solution of the LASSO Program

The Active Set

- First, assume that $\boldsymbol{Z}$ is such that the LASSO solution is unique.
- Let $A = \mathsf{supp}[\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)]$ be the LASSO active set (the nonzero coefficient estimates).
- Define $\boldsymbol{sg}_A = \mathsf{sign}[\widehat{\boldsymbol{\beta}}_{A,LASSO}(\lambda)]$ as the signs of the active (nonzero) coefficients.

Solution of the LASSO Program

The Active Set

- First, assume that $\boldsymbol{Z}$ is such that the LASSO solution is unique.
- Let $A = \mathsf{supp}[\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)]$ be the LASSO active set (the nonzero coefficient estimates).
- Define $\boldsymbol{sg}_A = \mathsf{sign}[\widehat{\boldsymbol{\beta}}_{A,LASSO}(\lambda)]$ as the signs of the active (nonzero) coefficients.
- Therefore,

$$\boxed{\begin{aligned}\widehat{\boldsymbol{\beta}}_{A,LASSO}(\lambda) &= \left(\boldsymbol{Z}'_A\boldsymbol{Z}_A\right)^{-1}\left(\boldsymbol{Z}'_A\boldsymbol{Y} - n\lambda\boldsymbol{sg}_A\right)\\ \widehat{\boldsymbol{\beta}}_{A^c,LASSO}(\lambda) &= \mathbf{0}.\end{aligned}}$$

Solution of the LASSO Program

The Active Set

- First, assume that $\boldsymbol{Z}$ is such that the LASSO solution is unique.
- Let $A = \mathsf{supp}[\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)]$ be the LASSO active set (the nonzero coefficient estimates).
- Define $\boldsymbol{sg}_A = \mathsf{sign}[\widehat{\boldsymbol{\beta}}_{A,LASSO}(\lambda)]$ as the signs of the active (nonzero) coefficients.
- Therefore,

$$\boxed{\begin{aligned}\widehat{\boldsymbol{\beta}}_{A,LASSO}(\lambda) &= \left(\boldsymbol{Z}'_A\boldsymbol{Z}_A\right)^{-1}\left(\boldsymbol{Z}'_A\boldsymbol{Y} - n\lambda\boldsymbol{sg}_A\right)\\ \widehat{\boldsymbol{\beta}}_{A^c,LASSO}(\lambda) &= \mathbf{0}.\end{aligned}}$$

- $\widehat{\boldsymbol{\beta}}_{A^c,LASSO}(\lambda)$ are the coefficient estimates that do not belong to the active set A.

Solution of the LASSO Program

The Active Set

- First, assume that $\boldsymbol{Z}$ is such that the LASSO solution is unique.
- Let $A = \mathsf{supp}[\widehat{\boldsymbol{\beta}}_{LASSO}(\lambda)]$ be the LASSO active set (the nonzero coefficient estimates).
- Define $\boldsymbol{sg}_A = \mathsf{sign}[\widehat{\boldsymbol{\beta}}_{A,LASSO}(\lambda)]$ as the signs of the active (nonzero) coefficients.
- Therefore,

$$\boxed{\begin{aligned}\widehat{\boldsymbol{\beta}}_{A,LASSO}(\lambda) &= \left(\boldsymbol{Z}'_A\boldsymbol{Z}_A\right)^{-1}\left(\boldsymbol{Z}'_A\boldsymbol{Y} - n\lambda\boldsymbol{sg}_A\right)\\ \widehat{\boldsymbol{\beta}}_{A^c,LASSO}(\lambda) &= \mathbf{0}.\end{aligned}}$$

- $\widehat{\boldsymbol{\beta}}_{A^c,LASSO}(\lambda)$ are the coefficient estimates that do not belong to the active set A.
- The coefficient estimates of the active set, $\widehat{\boldsymbol{\beta}}_{A,LASSO}(\lambda)$, are given by the OLS coefficients minus the amount $n\lambda(\boldsymbol{Z}'_A\boldsymbol{Z}_A)^{-1}\boldsymbol{sg}_A$. Important: $\boldsymbol{Z}'_A\boldsymbol{Z}_A$ is invertible.

Solution of the LASSO Program

The Active Set

- The shrinkage term $T\lambda(\boldsymbol{Z}'_A\boldsymbol{Z}_A)^{-1}\boldsymbol{sg}_A$ does not always move the OLS coefficients towards zero.

Solution of the LASSO Program

The Active Set

- The shrinkage term $T\lambda(\boldsymbol{Z}'_A\boldsymbol{Z}_A)^{-1}\boldsymbol{sg}_A$ does not always move the OLS coefficients towards zero.
- This happens due to correlation among active variables.

Solution of the LASSO Program

The Active Set

- The shrinkage term $T\lambda(\boldsymbol{Z}'_A\boldsymbol{Z}_A)^{-1}\boldsymbol{sg}_A$ does not always move the OLS coefficients towards zero.
- This happens due to correlation among active variables.
- However, the LASSO solution has a strictly smaller ℓ_1 norm than the OLS estimator on the active set:

$$\begin{aligned}\left\|\widehat{\boldsymbol{\beta}}_{A,LASSO}(\lambda)\right\|_1 &= \boldsymbol{sg}'_A(\boldsymbol{Z}'_A\boldsymbol{Z}_A)^{-1}\boldsymbol{Z}'_A\boldsymbol{Y} - n\lambda\boldsymbol{sg}'_A(\boldsymbol{Z}'_A\boldsymbol{Z}_A)^{-1}\boldsymbol{sg}_A \\ &< \left\|(\boldsymbol{Z}'_A\boldsymbol{Z}_A)^{-1}\boldsymbol{Z}'_A\boldsymbol{Y}\right\|_1 .\end{aligned}$$

Solution of the LASSO Program

Coordinate Descent (Shooting Algorithm)

1 Initialize $\boldsymbol{\beta} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda \boldsymbol{I})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$

2 Repeat

3 for $j = 1, \ldots, p$ do

4 $$r_{tj} = Y_i - \sum_{k=1, k\neq j}^{p} X_{tk}\beta_k$$

5 $$\beta_j = S_\lambda\left(\frac{1}{n}\boldsymbol{Z}_j'\boldsymbol{r}_j\right)$$

6 until convergence

Empirical Example: Sparse Signals in the Cross-Section of Returns

Chinco, Alex, Adam D. Clark-Joseph, and Mao Ye (2019). *Sparse Signals in the Cross-Section of Returns*. **Journal of Finance**, 74, 449–492.

Overview

Chinco, Clark-Joseph and Ye (JF, 2019)

- LASSO to make rolling 1-minute-ahead return forecasts using the entire cross-section of lagged returns as candidate predictors.

Overview

Chinco, Clark-Joseph and Ye (JF, 2019)

- LASSO to make rolling 1-minute-ahead return forecasts using the entire cross-section of lagged returns as candidate predictors.
- The LASSO increases both out-of-sample fit and forecast-implied Sharpe ratios. And, this out-of-sample success comes from identifying predictors that are unexpected, short-lived, and sparse.

Overview

Chinco, Clark-Joseph and Ye (JF, 2019)

- LASSO to make rolling 1-minute-ahead return forecasts using the entire cross-section of lagged returns as candidate predictors.
- The LASSO increases both out-of-sample fit and forecast-implied Sharpe ratios. And, this out-of-sample success comes from identifying predictors that are unexpected, short-lived, and sparse.
- Although the LASSO uses a statistical rule rather than economic intuition to identify predictors, the predictors it identifies are nevertheless associated with economically meaningful events: the LASSO tends to identify as predictors stocks with news about fundamentals.

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

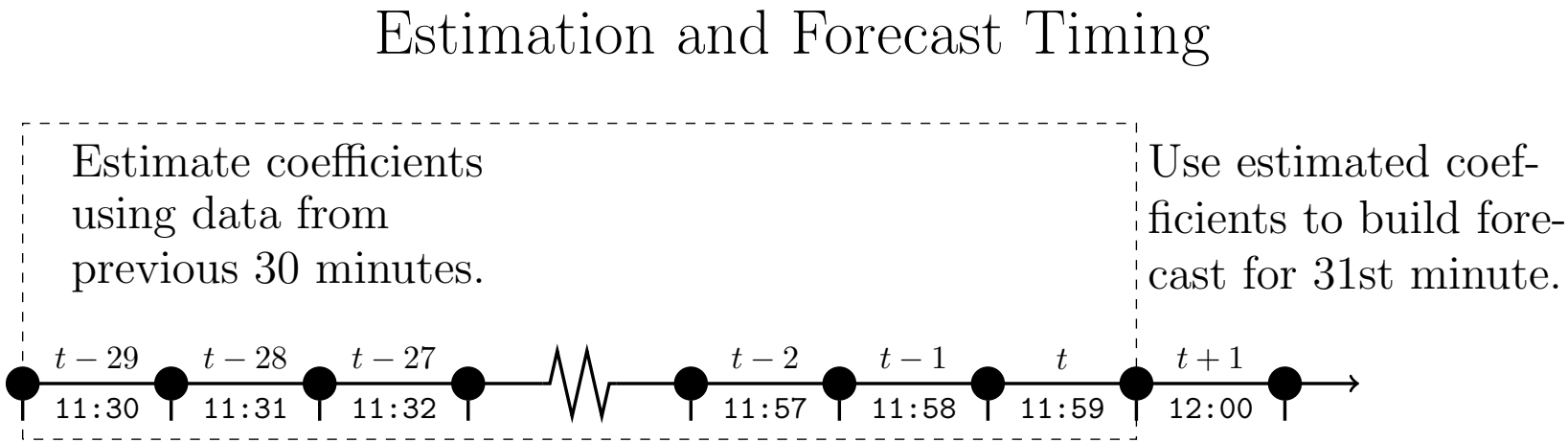


Figure 1: *To make our 1-minute-ahead forecast for the nth stock's return in minute* $(t+1) =$ 12:00*, we first estimate a model using data from the previous* 30 *minutes,* {11:30, . . . , 11:59}*. Then, we apply the estimated coefficients to the most recent* 3*-minutes of data,* {11:57, 11:58, 11:59}*. We use* $f^{\text{LASSO}}_{11:59}$ *to denote this forecast for the nth stock's return in minute* 12:00*, because it only uses information up to minute* 11:59*.*

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

Out-of-Sample Fit, The LASSO

	Mean	95% CI
$\bar{a}_n$ [%/m]	0.002 (0.002)	·
$\bar{b}_n$ [%/m]	1.433 (0.017)	[1.399, 1.467]
$\bar{R}^2_n$ [%]	2.467 (0.027)	[2.414, 2.520]

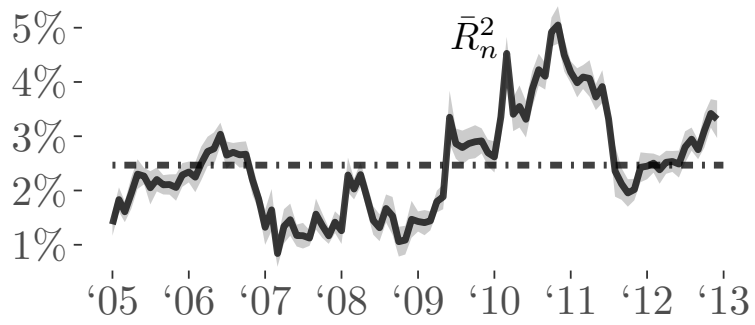


Table I: *The LASSO's out-of-sample fit.* **(Sample)** *Results of predictive regressions run at the stock-day level for* 250 *randomly chosen stocks on each trading day from January 2005 to December 2012.* **(Left)** *Population averages for regression coefficients and adjusted* R^2 *statistic. Numbers in parentheses are standard errors clustered by stock-day.* 95% CI *reports* 95%*-confidence intervals for population averages.* **(Right)** *Average adjusted* R^2 *statistic each month with the dashed line representing* $\bar{R}^2_n = 2.467\%$*. Grey bands denote the* 99.9% *confidence interval computed using standard errors clustered by stock-day.* **(Reads)** *"On average, the LASSO's* 1*-minute-ahead return forecast explains* 2.467% *of the variation in a stock's returns on a given day. And, the LASSO explains at least* 1% *of the variation in returns during every month in our sample."*

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

Increase in Out-of-Sample Fit, Main Results

	$\bar{R}^{2,\text{Bmk}}_n$ [%]	$\Delta\bar{R}^2_n$ [%pt]		p-val.
AR(3)	7.365 (0.076)	1.185 (0.012)	[1.162, 1.208]	0.000
Market	0.311 (0.003)	2.469 (0.027)	[2.416, 2.522]	0.000
AR(3), Market	5.553 (0.058)	1.424 (0.015)	[1.395, 1.453]	0.000
AR(1)	6.061 (0.052)	1.288 (0.013)	[1.263, 1.314]	0.000
AR(2)	7.309 (0.071)	1.165 (0.012)	[1.143, 1.188]	0.000
AR(4)	7.174 (0.076)	1.238 (0.012)	[1.214, 1.262]	0.000
AR(5)	6.725 (0.074)	1.307 (0.013)	[1.282, 1.332]	0.000
AR($h^\star$)	8.031 (0.080)	1.134 (0.011)	[1.113, 1.156]	0.000
Market, Industry	0.314 (0.007)	2.436 (0.027)	[2.384, 2.489]	0.000
Market, Size, Value	0.214 (0.003)	2.443 (0.027)	[2.390, 2.496]	0.000
AR(3), Market, Industry, Size, Value	1.443 (0.015)	2.232 (0.024)	[2.184, 2.279]	0.000

Table II: *The LASSO's increase in out-of-sample fit relative to a variety of benchmark models measured as the percentage point increase in adjusted* R^2. **(Sample)** *Regression results for a randomly selected subset of* 250 *stocks on each trading day from January 2005 to December 2012.* **(Estimates)** $\bar{R}^{2,\text{Bmk}}_n$*: Out-of-sample fit of a benchmark model measured using the adjusted* R^2 *statistic.* $\Delta\bar{R}^2_n$*: The LASSO's increase in out-of-sample fit.* p-val.*: Probability of observing the realized increase in out-of-sample fit* $\Delta\bar{R}^2_n$ *under null hypothesis of no increase. Numbers in parentheses are standard errors clustered by stock-day. Numbers in square brackets represent the* 95% *confidence interval for the mean.* **(Benchmark Models)** AR(3)*:* 3 *lags of a stock's own returns.* Market*:* 3 *lags of the market's returns.* AR(3), Market*:* 3 *lags of a stock's own returns and* 3 *lags of the market's returns.* AR(1)*:* 1 *lag of a stock's own returns.* AR(2)*:* 2 *lags of a stock's own returns.* AR(4)*:* 4 *lags of a stock's own returns.* AR(5)*:* 5 *lags of a stock's own returns.* AR($h^\star$)*:* $h^\star$ *lags of a stock's own returns where* $h^\star$ *is chosen within each* 30*-minute estimation window using the Bayesian information criteria.* Market, Industry*:* 3 *lags of the market's returns and* 3 *lags of the stock's industry's returns.* Market, Size, Value*:* 3 *lags of the market's returns, a size portfolio's returns, and a value portfolio's returns respectively.* AR(3), Market, Industry, Size, Value*:* 3 *lags of the market's returns, a stock's industry's returns, a size portfolio's returns, and a value portfolio's returns respectively.* **(Reads)** *"A researcher could explain an additional* $\Delta\bar{R}^2_n = 1.185\%pt$ *of the variation in returns by using both the LASSO and an* AR(3) *model rather than just the* AR(3) *model alone."*

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

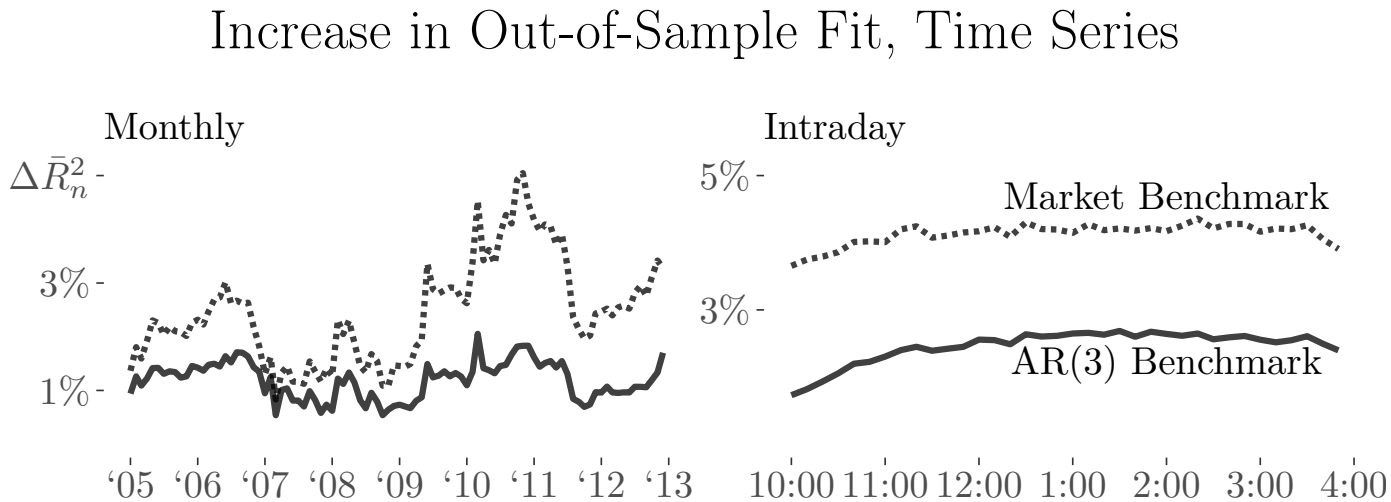


Figure 2: *Time-series variation in the LASSO's increase in out-of-sample fit relative to the* AR(3) *and market benchmarks. Increase in out-of-sample fit is measured as the percentage point increase in adjusted* R^2. **(Sample)** *Regression results for a randomly selected subset of* 250 *stocks on each trading day from January 2005 to December 2012.* **(Monthly)** *The LASSO's increase in out-of-sample fit by month.* **(Intraday)** *The LASSO's increase in out-of-sample fit by minute of the trading day.* **(Reads)** *"The LASSO increased in out-of-sample fit less during the financial crisis, but the LASSO's increase in out-of-sample fit is relatively stable over the course of the trading day."*

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

Increase in Out-of-Sample Fit, by Characteristics

AR(3) Benchmark	High, $>$ 50%ile $\Delta\bar{R}^2_n$ [%pt]		p-val.	Low, $<$ 50%ile $\Delta\bar{R}^2_n$ [%pt]		p-val.
Mkt Cap	1.341 (0.017)	[1.307, 1.375]	0.000	1.029 (0.011)	[1.007, 1.051]	0.000
Volume	1.347 (0.017)	[1.315, 1.380]	0.000	1.023 (0.011)	[1.002, 1.044]	0.000
Volatility	1.137 (0.012)	[1.113, 1.162]	0.000	1.234 (0.016)	[1.204, 1.265]	0.000
Spread	0.996 (0.011)	[0.975, 1.017]	0.000	1.377 (0.017)	[1.343, 1.410]	0.000

Market Benchmark	High, $>$ 50%ile $\Delta\bar{R}^2_n$ [%pt]		p-val.	Low, $<$ 50%ile $\Delta\bar{R}^2_n$ [%pt]		p-val.
Mkt Cap	2.605 (0.036)	[2.534, 2.675]	0.000	2.332 (0.027)	[2.280, 2.385]	0.000
Volume	2.518 (0.033)	[2.454, 2.582]	0.000	2.419 (0.030)	[2.360, 2.478]	0.000
Volatility	2.680 (0.028)	[2.626, 2.734]	0.000	2.259 (0.032)	[2.196, 2.321]	0.000
Spread	2.302 (0.026)	[2.251, 2.353]	0.000	2.639 (0.036)	[2.569, 2.709]	0.000

Table III: *The LASSO's increase in out-of-sample fit relative to the* AR(3) *and market benchmarks sorted by firm characteristics. Increase in out-of-sample fit is measured as the percentage point increase in adjusted* R^2. **(Sample)** *Regression results for a randomly selected subset of* 250 *stocks on each trading day from January 2005 to December 2012.* **(Estimates)** $\Delta\bar{R}^2_n$*: The LASSO's increase in out-of-sample fit relative a benchmark model.* p-val.*: Probability of observing the realized* $\Delta\bar{R}^2_n$ *under the null hypothesis of no increase. Numbers in parentheses are standard errors clustered by stock-day. Numbers in square brackets are* 95% *confidence intervals.* **(Characteristics)** Mkt Cap*: Market value at close of previous trading day.* Volume*: Trading volume on previous trading day.* Volatility*: Volatility of* 1*-minute returns during previous trading day.* Spread*: Average bid-ask spread during previous trading day.* High, $>$ 50%ile*: Subset of stocks with above-median values for a given characteristic.* Low, $<$ 50%ile*: Subset of stocks with below-median values for the same characteristic.* **(Reads)** *"The LASSO increases out-of-sample fit slightly more for large, liquid, frequently-traded stocks."*

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

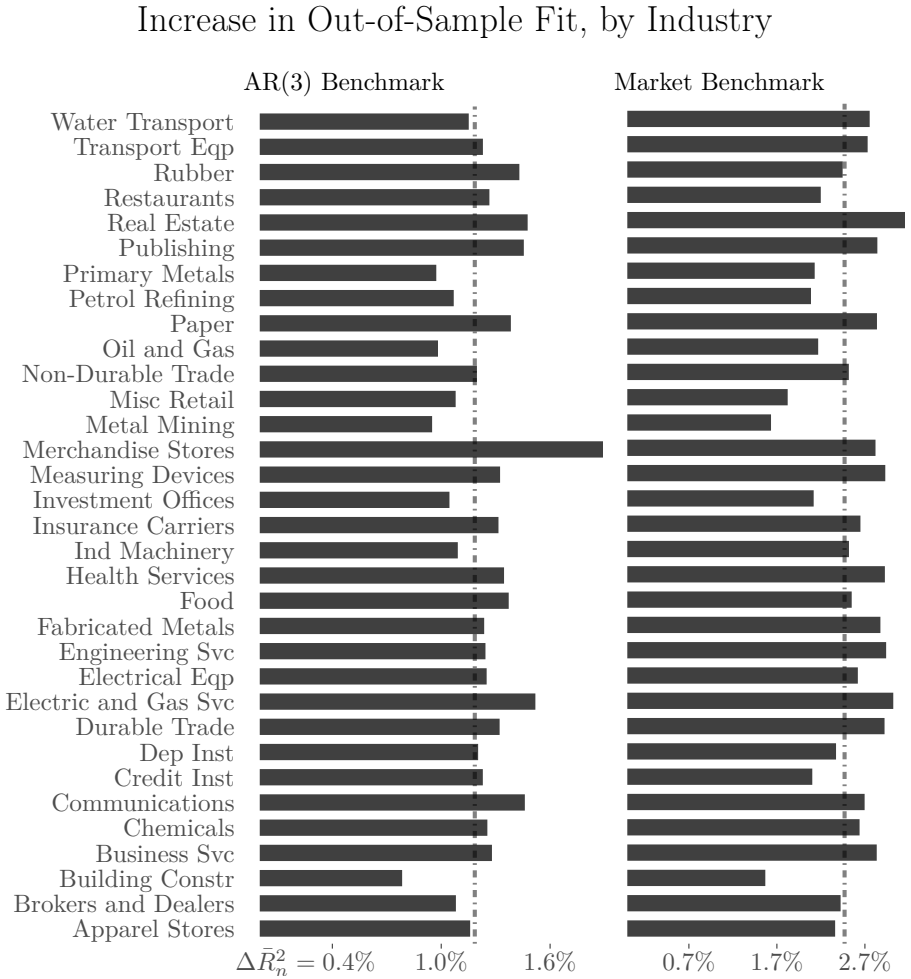


Figure 3: *The LASSO's increase in out-of-sample fit relative to the* AR(3) *and market benchmarks sorted by* 3*-digit SIC industries. Increase in out-of-sample fit is measured as the percentage point increase in adjusted* R^2*.* **(Sample)** *Regression results for a randomly selected subset of* 250 *stocks on each trading day from January 2005 to December 2012. We restrict the sample to industries with at least* 20 *firms on average.* **(Left)** *The LASSO's increase in out-of-sample fit relative to a* AR(3) *benchmark with the dashed line representing* $\Delta\bar{R}^2_n = 1.185\%$pt*.* **(Right)** *The LASSO's increase in out-of-sample fit relative to a market benchmark with the dashed line representing* $\Delta\bar{R}^2_n = 2.469\%$pt*.* **(Reads)** *"The LASSO increases out-of-sample fit for all industries."*

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

Forecast-Implied Performance Net of Trading Costs

Annualized Sharpe Ratios

S&P 500	LASSO	AR(3)
0.123	1.791	−0.662

LASSO-Implied Strategy Abnormal Returns [%/yr]	α	Mkt	HmL	SmB	Mom
Market	2.709 (0.034)	0.004 (0.002)			
3-Factor Model	2.713 (0.034)	0.004 (0.002)	−0.004 (0.004)	0.000 (0.003)	
4-Factor Model	2.707 (0.034)	0.005 (0.002)	−0.004 (0.004)	0.003 (0.004)	0.003 (0.004)

Table IV: *Performance of forecast-implied strategies net of trading costs.* **(Sample)** *Each trading day from January 2005 to December 2012.* **(Sharpe Ratios)** *Annualized Sharpe ratios of forecast-implied trading strategies net of trading costs. First column reports results for strategy that invests* \$1 *in the S&P 500 at market open on January 3rd, 2005 and holds that position until market close on December 31st, 2012. The next column reports results for the LASSO's forecast-implied trading strategy over the same time period. The third column reports analogous results for an* AR(3)*-implied strategy with the same initial investment.* **(Abnormal Returns)** *Net abnormal returns of the LASSO-implied strategy relative to the market, the Fama and French (1993)* 3*-factor model, and the Carhart (1997)* 4*-factor model. The size of the initial investment in the LASSO-implied strategy was chosen so that it has the same average excess return as the buy-and-hold S&P 500 strategy. First column reports annualized abnormal returns. Remaining columns report dimensionless slope coefficients associated with each factors.* **(Reads)** *"The LASSO-implied strategy generates positive excess returns net of the spread with an annualized Sharpe ratio of* 1.791*, and these excess returns are not explained by the strategy's exposures to standard risk factors."*

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

Trading Frequency of Forecast-Implied Strategies

Aggregate [#/min]	LASSO	AR(3)
Trades	8.624	17.643
Buy Orders	4.306	8.887
Successful	2.600	6.538

By Characteristics [#/min]	>50%ile Mkt Cap	>50%ile Volume	>50%ile Volatility	<50%ile Spread
LASSO Trades	5.188	5.115	4.821	5.114
LASSO Buy Orders	2.592	2.555	2.407	2.556
LASSO Successful	1.578	1.533	1.467	1.550

Table V: *Trading frequency of forecast-implied strategies.* **(Sample)** *Minute-level trades for randomly selected subset of* 250 *stocks on each trading day from January 2005 to December 2012 for which we compute* 1*-minute-ahead return forecasts using the LASSO.* **(Aggregate)** *First row reports the average number of trades per minute made by LASSO- and* AR(3)*-implied strategies. Second row reports the number of buy orders per minute. And, third row reports the number of trades per minute that made money after accounting for the bid-ask spread.* **(By Characteristics)** *Number of trades, buy orders, and successful trades per minute for the LASSO-implied strategy among large stocks, stocks with high trading volume, stocks with high return volatility, and liquid stocks.* **(Reads)** *"The* AR(3)*-implied strategy trades roughly twice as often as the LASSO-implied strategy. And, the LASSO-implied strategy is more active in large, liquid, frequently traded stocks."*

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

Characteristics of LASSO Predictors

	Dep. Variable: Ever Selected					
Mkt Cap > 50%ile	0.005 (0.004) [0.11%]					0.003 (0.007) [0.07%]
Volume > 50%ile		0.004 (0.004) [0.09%]				0.001 (0.006) [0.01%]
Volatility > 50%ile			−0.005 (0.004) [0.13%]			−0.004 (0.004) [0.10%]
Spread < 50%ile				0.002 (0.004) [0.05%]		0.006 (0.006) [0.15%]
In Same Industry					−0.046 (0.006) [1.15%]	−0.046 (0.006) [1.14%]

Table VI: *Characteristics of the predictors selected by the LASSO.* **(Sample)** *On each trading day, d, from January 2005 to December 2012 the data contain one observation per predictor for each of the randomly selected* 250 *stocks for which we make* 1*-minute-ahead return forecasts. For example, on October 6th, 2010 the LASSO could choose as predictors any of the* $N = 2{,}191$ *NYSE-listed stocks. So, on that day, our data contained* $250 \times 2{,}191 = 547{,}750$ *separate observations.* **(Specification)** *Each column reports the results from a different logit regression. Point estimates are log-odds ratios. Numbers in parentheses are standard errors clustered by day and predictor (i.e., stock* n'*). Numbers in square brackets are the marginal effects implied by the log-odds ratios.* **(Variables)** *Dep. Variable: An indicator variable that is one if stock* n' *was ever used by the LASSO as a predictor when making* 1*-minute-ahead return forecasts for stock* n *on day* d*.* Mkt Cap > 50%ile*: An indicator variable that is one if stock* n' *had above-median market capitalization on day* d*.* Volume > 50%ile*: An indicator variable that is one if stock* n' *had above-median trading volume on day* d*.* Volatility > 50%ile*: An indicator variable that is one if stock* n' *had above-median* 1*-minute-return volume on day* d*.* Spread < 50%ile*: An indicator variable that is one if stock* n' *had a below-median average bid-ask spread on day* d*.* In Same Industry*: An indicator variable that is one if stock* n *and stock* n' *both belonged to the same* 3*-digit SIC industry.* **(Reads)** *"There is little correlation between the LASSO's choice of predictors and firm size, trading volume, return volatility, and bid-ask spread. And, the LASSO is less likely to select a stock from the same industry as a predictor."*

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

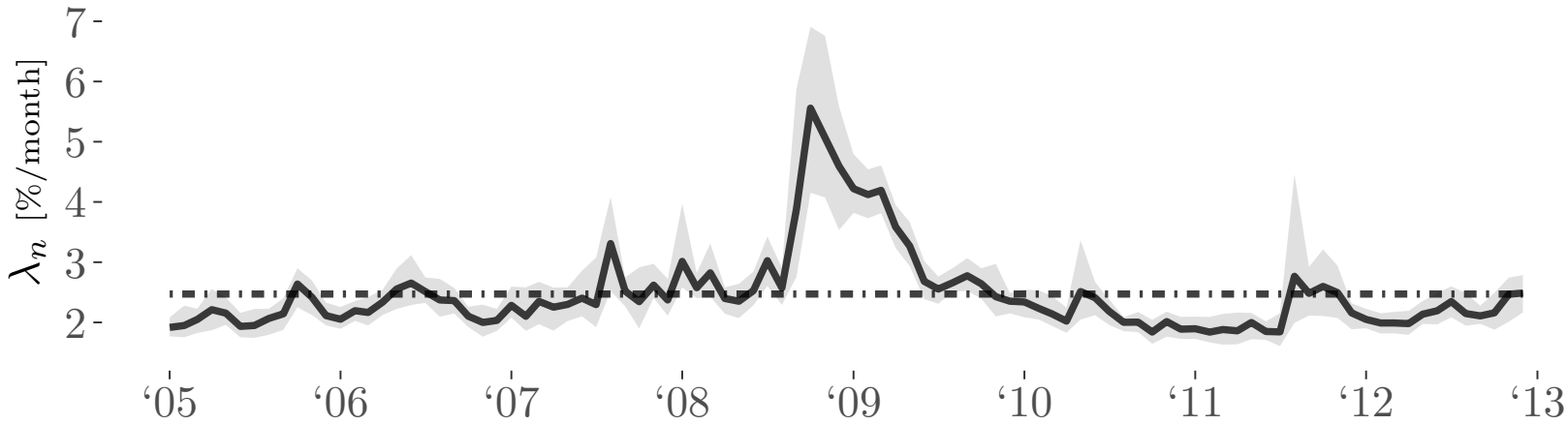


Figure 4: *Monthly average (solid line) and* 95% *confidence interval (gray bands) for the LASSO's cross-validated penalty parameter,* λ_n*, in units of* % *per month. Dashed line denotes sample average over the course of the entire sample period, which is* 2.5%. **(Sample)** *Estimation results for a randomly selected subset of* 250 *stocks on each trading day from January 2005 to December 2012.* **(Reads)** *"The LASSO typically ignores all predictors weaker than* $\lambda = 2.5\%$ *per month when making its* 1*-minute-ahead return forecasts."*

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

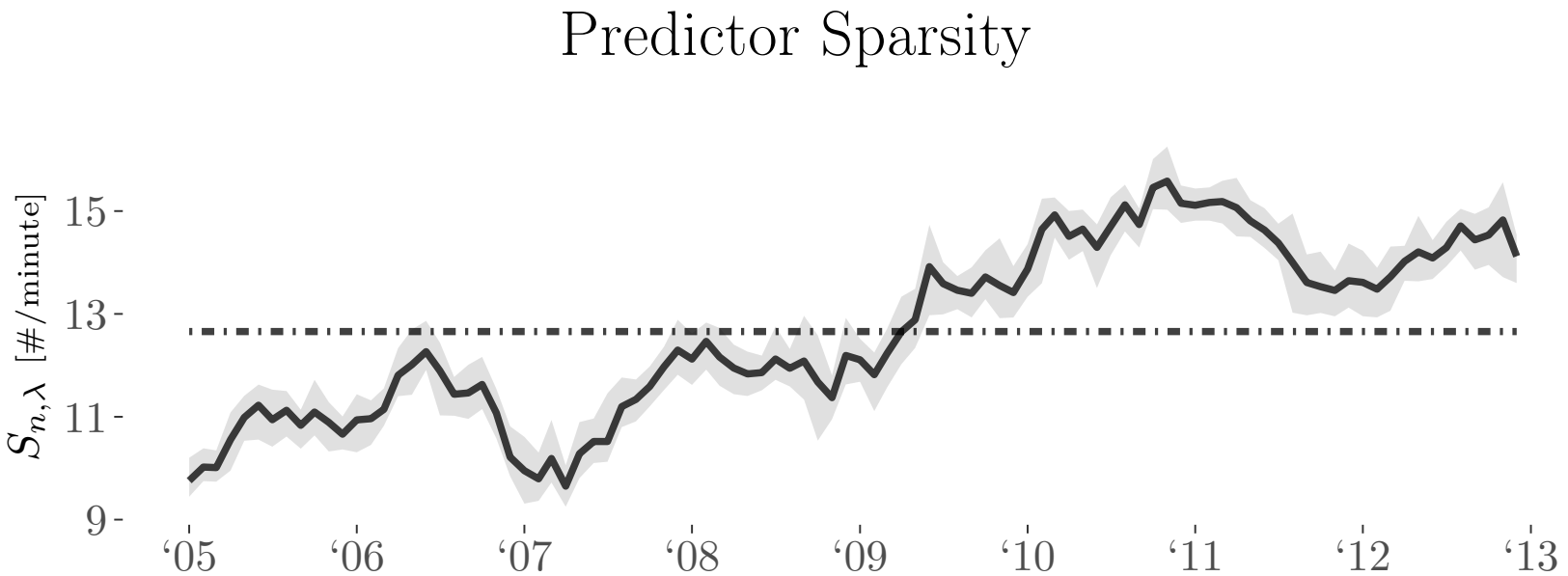


Figure 6: *Average (solid line) and* 95% *confidence interval (gray bands) for the number of predictors,* $S_{n,\lambda}$*, used by the LASSO for each of its* 1*-minute-ahead return forecasts in a given month. Dashed line denotes sample average over the course of the entire sample period, which is* 12.7 *predictors.* **(Sample)** *Estimation results for a randomly selected subset of* 250 *stocks on each trading day from January 2005 to December 2012.* **(Reads)** *"On average, the LASSO uses only* 12.7 *predictors (out of a possible* $3 \cdot N \approx$ 6,000*) when making its* 1*-minute-ahead return forecasts for each stock."*

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

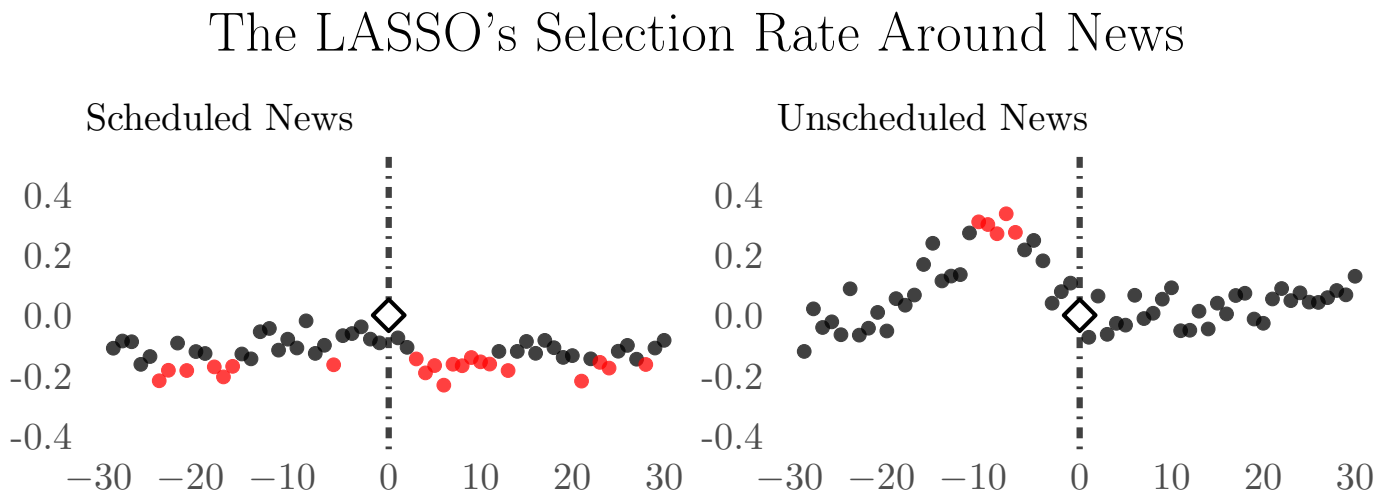


Figure 7: *x-axis: event time relative to news flash (in minutes). Vertical dashed line: minute of news flash about n'th stock's revenues, $h = 0$. y-axis: difference between the number of times that the LASSO selected the n'th stock as a predictor when making its* 1*-minute-ahead return forecast for minute h and the number of times it did so for minute $h = 0$. Each dot denotes the difference in the LASSO's selection rate in minute h relative to minute $h = 0$ with the large diamond denoting a difference of zero for $h = 0$ (a normalization). Red dots denote differences that are statistically significant at the* 5% *level using standard errors clustered by year.* **(Scheduled News)** *News flashes about scheduled events.* **(Unscheduled News)** *News flashes about unscheduled events.* **(Sample)** 61*-minute window around each novel news flash about an NYSE-listed stock's revenues which occurred during normal trading hours from January 2005 to December 2012.* **(Reads)** *"When there is a scheduled news flash about the n'th stock's revenues in minute t, the LASSO selects stock n' as a predictor slightly more often when making its* 1*-minute-ahead return forecasts for minute t—that is, for $h = 0$. But, if the news flash is unscheduled, then the LASSO selects stock n' as a predictor much more often when making its* 1*-minute-ahead return forecasts in the* 10 *minutes prior to minute t."*

Results

Chinco, Clark-Joseph and Ye (JF, 2019)

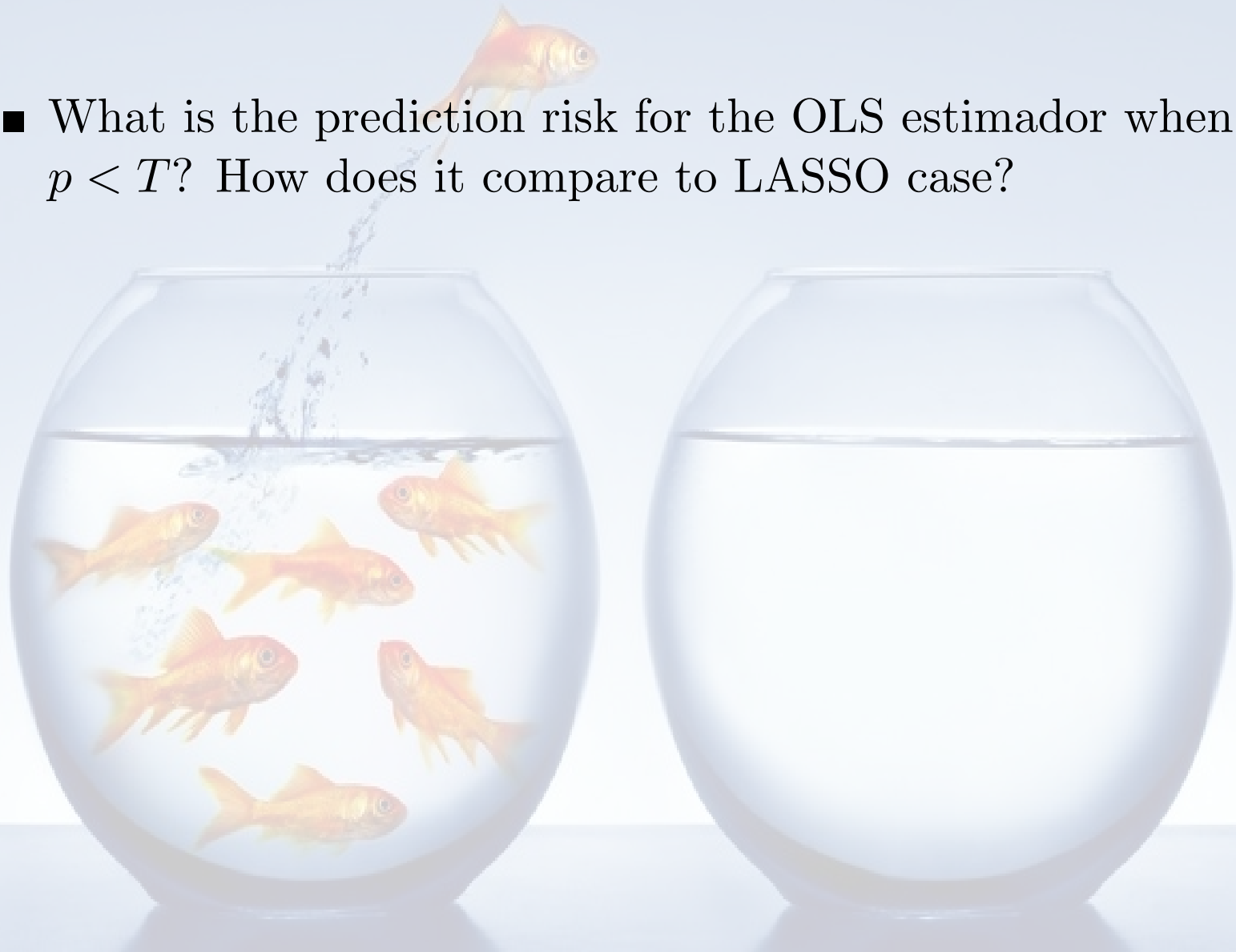
The LASSO's Selection Rate Around News

	Dep. Variable: Times Selected [#/min]					
	Scheduled News			Unscheduled News		
$1_{\{-30\leq h<0\}}$	−0.115 (0.058)			0.113 (0.101)		
$1_{\{h=0\}}$	·			·		
$1_{\{30\geq h>0\}}$	−0.141 (0.054)			0.026 (0.091)		
$1_{\{-30\leq h<-20\}}$		−0.139 (0.072)	−0.139 (0.073)		−0.029 (0.120)	−0.029 (0.120)
$1_{\{-20\leq h<-10\}}$		−0.124 (0.070)	−0.123 (0.070)		0.153 (0.124)	0.153 (0.124)
$1_{\{-10\leq h<0\}}$		−0.086 (0.056)	−0.085 (0.056)		0.203 (0.094)	0.202 (0.094)
$1_{\{h=0\}}$		·	·		·	·
$1_{\{10\geq h>0\}}$		−0.156 (0.050)	−0.157 (0.050)		0.008 (0.086)	0.008 (0.086)
$1_{\{20\geq h>10\}}$		−0.126 (0.063)	−0.127 (0.063)		0.001 (0.108)	0.001 (0.108)
$1_{\{30\geq h>20\}}$		−0.142 (0.066)	−0.142 (0.066)		0.069 (0.109)	0.069 (0.109)
Impact			0.201 (0.318)			−0.142 (0.876)
Sentiment			0.170 (0.209)			−1.090 (0.659)
Year FE	Y	Y	Y	Y	Y	Y

Table VII: *Change in the LASSO's selection rate around news flashes about a company's revenues.* **(Specification)** *Each column reports the results from a different regression. Point estimates have units of number of times selected per minute. Numbers in parentheses are standard errors clustered by year. Dots denote omitted estimates for the reference category,* $h = 0$. **(Variables)** *Dep. Variable: Number of times that the* n'*th stock was selected as a predictor by the LASSO in minute* $(t + h)$. Impact*: A zero (low) to one (high) variable summarizing news flash's impact on overall market volatility during the next several hours.* Sentiment*: A zero (very negative) to one (very positive) variable summarizing the sentiment of the text contained in a news flash.* **(Scheduled News)** *News flashes about scheduled events.***(Unscheduled News)** *News flashes about unscheduled events.* **(Sample)** 61*-minute window around each novel news flash about an NYSE-listed stock's revenues which occurred during normal trading hours from January 2005 to December 2012.* **(Reads)** *"The LASSO uses stock* n' *to make* 0.115 *fewer* 1*-minute-ahead return forecasts in each of the* 30 *minutes leading up to a scheduled news flash,* $1_{-30\leq h<0}$*, than it does in the minute of the news flash itself,* $h = 0$*. By contrast, the LASSO uses stock* n' *to make* 0.203 *more* 1*-minute-ahead return forecasts in each of the* 10 *minutes leading up to an unscheduled news flash,* $1_{-10\leq h<0}$*."*

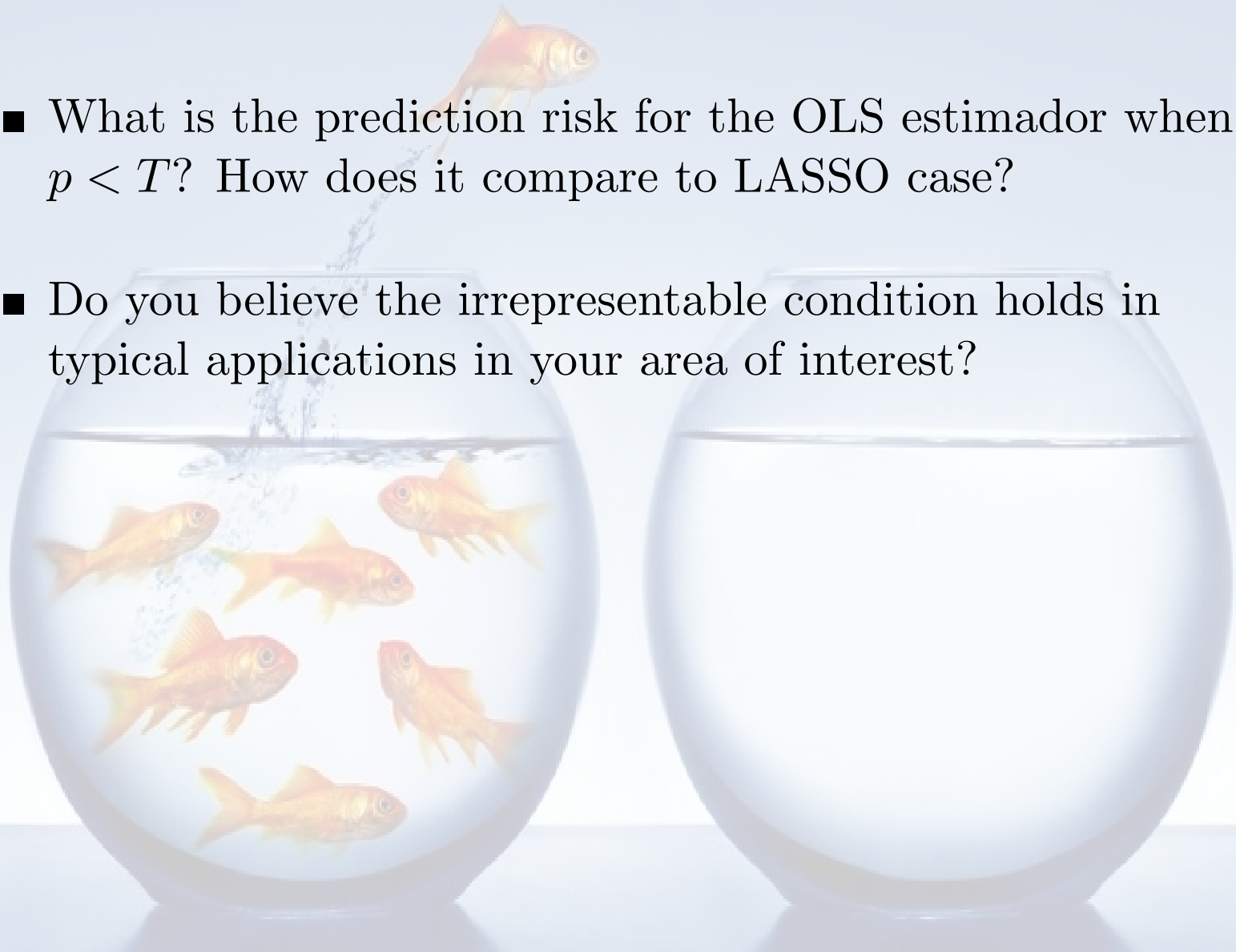
Motivating Questions

- What is the prediction risk for the OLS estimador when $p < T$? How does it compare to LASSO case?



Motivating Questions

- What is the prediction risk for the OLS estimador when $p < T$? How does it compare to LASSO case?
- Do you believe the irrepresentable condition holds in typical applications in your area of interest?



Motivating Questions

- What is the prediction risk for the OLS estimador when $p < T$? How does it compare to LASSO case?
- Do you believe the irrepresentable condition holds in typical applications in your area of interest?
- Think about an application of what you have learned today to the industry of your interest.