Linear Model in High Dimensions, II: Estimation and Inference

Jesper Riis–Vestergaard Sørensen

University of Copenhagen, Department of Economics

October 14, 2022

Recap

Last time:

High-dimensional framework:

$$p = p_n \text{ with } p/n \to \text{const.} > 0 \text{ as } n \to \infty.$$

- Allows 'wide' data sets (p/n not ≈ 0).

OLS poorly behaved in high dimensions ($p/n \nrightarrow 0$).

Introduced sparsity and Lasso.

Talked about tuning penalty selection.

... and implementation in Python.

Overview

- Estimation Error Control
  - Least Squares
  - Lasso
- Inference
  - Post-Double Lasso
  - Orthogonalized Moments
- Other Methods for High-Dimensional Regression
  - Dantzig Selector
  - Ridge Regression
  - Elastic Net

Estimation Error Control

Least Squares

Consistency in Low Dimensions, I

Linear mean regression model:

$$Y = \sum_{j=1}^{p} \beta_j X_j + \varepsilon = X'\beta + \varepsilon, \quad \mathrm{E}[\varepsilon | X] = 0.$$

Least squares (LS) estimator:

$$\widehat{\beta}^{\mathrm{LS}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Y}.$$

Low-dimensional regime (p fixed).

Consistency conditions?

Consistency in Low Dimensions, II

Main Conditions

$$\widehat{\beta}^{\text{LS}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}, \qquad \text{(Estimator)}$$

$$\Rightarrow \widehat{\beta}^{\text{LS}} - \beta = \left(\mathbf{X}'\mathbf{X}/n\right)^{-1}\left(\mathbf{X}'\varepsilon/n\right). \qquad \text{(Estimation Error)}$$

Consistency follows from two conditions + Slutsky:

1. $\mathbf{X}'\mathbf{X}/n \to_{\mathrm{P}}$ to nonsingular matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$.
    - In 1D: Just ruling out division by zero.
2. $\mathbf{X}'\varepsilon/n \to_{\mathrm{P}}$ to zero vector $\mathbf{0} \in \mathbb{R}^{p}$.

Then $(\mathbf{X}'\mathbf{X}/n)^{-1}(\mathbf{X}'\varepsilon/n) \to_{\mathrm{P}} \mathbf{A}^{-1}\cdot\mathbf{0} = \mathbf{0}$.

Consistency in Low Dimensions, III

Singularity, Definiteness and Eigenvalues

For $\mathbf{M}$ positive semidefinite (p.s.d.),

$\mathbf{M}$ invertible $\Leftrightarrow$ $\mathbf{M}$ positive definite (p.d.) $\Leftrightarrow$ all positive eigenvalues.

Let $\Lambda_{\min}(\mathbf{M})$ = smallest eigenvalue of $\mathbf{M}$.

By CMT, '$\mathbf{X}'\mathbf{X}/n \to_{\mathrm{P}} \mathbf{A}$ nonsingular' means

$$\Lambda_{\min}(\mathbf{X}'\mathbf{X}/n) \xrightarrow{\mathrm{P}} \text{const.} > 0.$$

Error Bound in Low Dimensions

Estimation error:

$$\widehat{\beta}^{\text{LS}} - \beta = \left(\mathbf{X}'\mathbf{X}/n\right)^{-1}\left(\mathbf{X}'\varepsilon/n\right). \qquad \text{(in } \mathbb{R}^p\text{)}$$

In ℓ^2 (Euclidean) norm:

$$\|\widehat{\beta}^{\text{LS}} - \beta\|_2 = \|\left(\mathbf{X}'\mathbf{X}/n\right)^{-1}\left(\mathbf{X}'\varepsilon/n\right)\|_2. \qquad \text{(in } \mathbb{R}\text{)}$$

Linear algebra [skipped] shows error bound:

$$\|\widehat{\beta}^{\text{LS}} - \beta\|_2 \leqslant \frac{\|\mathbf{X}'\varepsilon/n\|_2}{\Lambda_{\min}(\mathbf{X}'\mathbf{X}/n)}.$$

Impossibility of OLS with $p > n$

$$\widehat{\beta}^{\text{LS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Inversion not possible when $p > n$...

Lemma

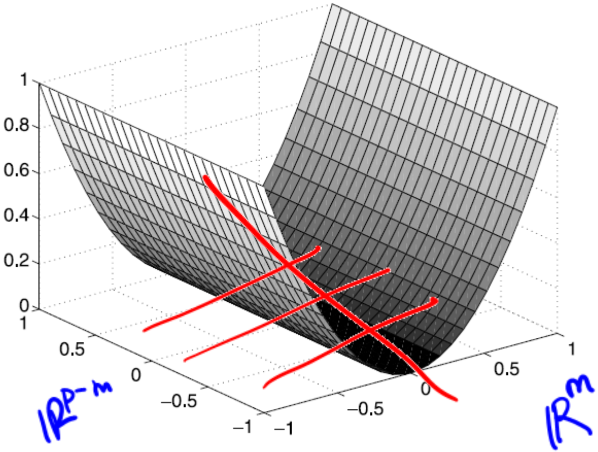
If $p > n$, then $\mathbf{X}'\mathbf{X}$ is (always) singular.

- RHS variables must be perfectly colinear in sample.

Proof: $\operatorname{rank}(\mathbf{X}'\mathbf{X}) = \operatorname{rank}(\mathbf{X}) \leqslant \min(n, p)$

Illustration of Impossibility of Least Squares

Figure: Sum of squares function in $p > n$ setting



Always flat in some direction.

Lasso

Consistency in High Dimensions, I

$$\text{Lasso:} \quad \widehat{\beta}(\lambda) \in \operatorname*{argmin}_{b \in \mathbf{R}^p} \Bigg\{ \underbrace{\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2}_{\text{(mis)fit}} + \underbrace{\lambda \|b\|_1}_{\text{penalty}} \Bigg\},$$

Penalty level $\lambda \geqslant 0$ of our choosing.

High-dimensional regime: $p/n \to$ const. > 0 as $n \to \infty$.

Consistency? Error bounds?

Consistency in High Dimensions, II

Conditions for Lasso analogous to LS

1. Want $\mathbf{X}'\varepsilon/n$ 'small'

2. Want $\mathbf{X}'\mathbf{X}/n$ 'well behaved'

RE 1: We will *choose* λ to force $\mathbf{X}'\varepsilon/n$ 'small.'

RE 2: Smallest eigenvalue of $\mathbf{X}'\mathbf{X}/n$ may be zero,

... but may *hope* small submatrices have nonzero eigenvalues.

Consistency in High Dimensions, III

Let $\mathbf{X}_J$ be submatrix of $\mathbf{X}$ with $\emptyset \neq J \subseteq \{1, 2, \ldots, p\}$ columns.

Recall $s = \sum_{j=1}^{p} \mathbf{1}\{\beta_j \neq 0\}$.

Smallest (s-)sparse eigenvalue,

$$\phi_{\min}(s) := \phi_{\min}(s)(\mathbf{X}'\mathbf{X}/n) := \min_{1 \leqslant |J| \leqslant s} \Lambda_{\min}(\mathbf{X}_J'\mathbf{X}_J/n).$$

Lasso only relies on invertibility of small submatrices

... OLS needs full invertibility.

Lasso Error Guarantees

Theorem

Let $c > 1$. *Then* $\lambda \geqslant c \max_{1 \leqslant j \leqslant p} |n^{-1} \sum_{i=1}^{n} \varepsilon_i X_{ij}|$ *implies*

$$\|\widehat{\beta}(\lambda) - \beta\|_2 \leqslant \text{const.}(c) \times \frac{\lambda\sqrt{s}}{\phi_{\min}(s)}.$$

[Proof: Skipped.]

Digest

$$\lambda \geqslant c \max_{1\leqslant j\leqslant p} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_{ij} \right|, \qquad \text{(Qualifier)}$$

$$\Rightarrow \|\widehat{\beta}(\lambda) - \beta\|_{\cancel{1}\,2} \leqslant \text{const.}(c) \times \frac{\lambda\sqrt{s}}{\phi_{\min}(s)}. \qquad \text{(Error Bound)}$$

Nonasymptotic: Holds for finite n and p.

Conditional: Qualifier suggests penalty (BRT rule...)

Trade-off: Want good bound ($\lambda \downarrow$) with high probability ($\lambda \uparrow$).

Bickel-Ritov-Tsybakov Rule, Again

Lemma

Let $\varepsilon \sim N(0, \sigma^2)$ be independent of X and $\lambda = \widehat{\lambda}^{\mathrm{BRT}}$ chosen according to the Bickel-Ritov-Tsybakov rule,

$$\widehat{\lambda}^{\mathrm{BRT}} = \frac{2c\sigma}{\sqrt{n}} \Phi^{-1}\left(1 - \frac{\alpha}{2p}\right) \max_{1 \leqslant j \leqslant p} \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_{ij}^2}.$$

Then $\lambda \geqslant c \max_{1 \leqslant j \leqslant p} |n^{-1} \sum_{i=1}^{n} \varepsilon_i X_{ij}|$ with probability at least $1 - \alpha$.

Moreover, λ satisfies upper bound

$$\widehat{\lambda}^{\mathrm{BRT}} \leqslant 2c\sigma \sqrt{\frac{2 \ln(2p/\alpha)}{n}} \max_{1 \leqslant j \leqslant p} \sqrt{\frac{1}{n} \sum_{i=1}^{n} X_{ij}^2}.$$

[Proof: Skipped.]

High-Probability Lasso Error Bound

Combine theorem and lemma: If

- errors are independent normal,
- BRT penalty, $\lambda = \widehat{\lambda}^{\texttt{BRT}}$,

then with probability at least $1 - \alpha$, have error bound

$$\|\widehat{\beta}(\lambda) - \beta\|_2 \leqslant C\sqrt{\frac{s \ln p}{n}}.$$

for some constant $C > 0$.

Lasso Consistency

If $\alpha = \alpha_n \to 0$, then error bnd holds with prob. approaching one.

Consistency follows if $(s/n)(\ln p) \to 0$.

Much weaker than $p/n \to 0$.

p may be much (e.g. exponentially) larger than n.

Extensions:

BCCK rule imply similar results w/o normality/homosked.

Chetverikov & Sørensen [2021] go beyond linear model.

Inference

Motivation

Suppose regressors $X = (D, Z')'$, where

- D: Variable of interest ('treatment').
- Z: Vector of controls. Possibly very long.

Model still

$$Y = \alpha_0 D + Z'\gamma_0 + \varepsilon, \quad \mathrm{E}[\varepsilon \mid D, Z] = 0.$$

Object of interest: α_0 (low-dimensional)

Q: How to construct confidence interval?

Lasso?

One possibility: Plain Lasso

1. Lasso Y_i using D_i and Z_i.

Yields $\widehat{\alpha}$ and $\widehat{\gamma}$ (for appropriate penalty).

Idea: Base CI on $\widehat{\alpha}$.

Lasso?

Issues:

1. $\widehat{\alpha}$ not analytically available, $\widehat{\alpha} = ?$

2. Exact distribution unknown/complicated, $\widehat{\alpha} \overset{d}{=} ?$

   - Orthonormal case: $\widehat{\alpha} = \text{sgn}(\widehat{\alpha}^{\text{LS}}) \left(|\widehat{\alpha}^{\text{LS}}| - \frac{\lambda}{2}\right)_+$

3. Asymptotic distribution unknown: $\sqrt{n}(\widehat{\alpha} - \alpha_0) \xrightarrow{d} ?$

$\Rightarrow$ No good approximation: $\widehat{\alpha} \overset{d}{\approx} ?$

$\Rightarrow$ Difficult to construct CI

Post-Lasso?

Another possibility:

1. Lasso Y_i using D_i and $Z_i \Rightarrow \widehat{\alpha}$ and $\widehat{\gamma}$
   - Gather *selection* $\widehat{J} := \{j; \widehat{\gamma}_j \neq 0\}$.
2. THEN: Least squares Y_i using D_i and $Z_{i\widehat{J}} \Rightarrow \widetilde{\alpha}$

Called Post-(Single)Lasso.

Q: Distribution?

REF: Belloni, Chernozhukov [2013 Bernoulli] "Least squares after model selection in high-dimensional sparse models."

Post-Lasso?

Simulations

Plotting

$$\frac{\sqrt{n}\,(\tilde{\alpha}-\alpha_0)}{\sqrt{\widehat{var}(\tilde{\alpha})}}$$

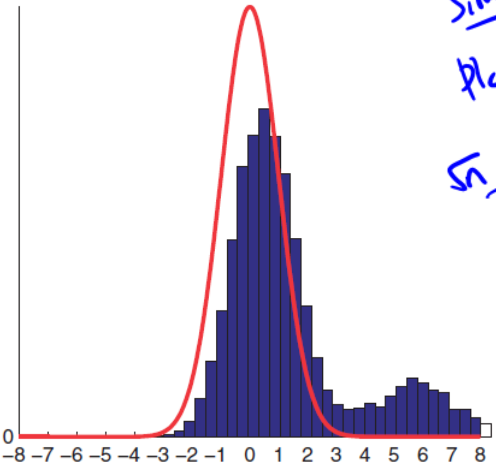


Figure: Post-Lasso (Normalized) vs. Standard Normal

Post-Lasso?

What went wrong?

- Refitting after Lasso selection.
- Relies on (unrealistic) perfect model selection.
- Very sensitive to mistakes.
- Omission of relevant control $\Rightarrow$ bias.

Post-Double Lasso

Strategy

Augment

$$Y = \alpha_0 D + Z'\gamma_0 + \varepsilon, \quad \mathrm{E}[\varepsilon \mid D, Z] = 0,$$

$\Rightarrow \mathrm{E}[\varepsilon f(D,Z)] = 0$

with 'first stage'

$$D = Z'\psi_0 + \nu, \quad \mathrm{E}[\nu \mid Z] = 0.$$

Added structure implies moment condition

$$\mathrm{E}\left[\left(D - Z'\psi_0\right)\underbrace{\left(Y - \alpha_0 D - Z'\gamma_0\right)}_{=\varepsilon}\right] = 0.$$

Hence

$$\alpha_0 = \frac{\mathrm{E}\left[\left(D - Z'\psi_0\right)\left(Y - Z'\gamma_0\right)\right]}{\mathrm{E}\left[\left(D - Z'\psi_0\right)D\right]}.$$

Suggests strategy.

Construction

Post-Double Lasso consists of three steps:

1. Lasso D_i using $Z_i \Rightarrow \widehat{\psi}$
2. Lasso Y_i using D_i and $Z_i \Rightarrow \widehat{\alpha}$ and $\widehat{\gamma}$
3. Estimate α_0 per analogy principle:

$$\check{\alpha} := \frac{\sum_{i=1}^{n}(D_i - Z_i'\widehat{\psi})(Y_i - Z_i'\widehat{\gamma})}{\sum_{i=1}^{n}(D_i - Z_i'\widehat{\psi})D_i}.$$

Result

Under (sparsity+) conditions, Post-Double Lasso satisfies

$$\frac{\sqrt{n}(\check{\alpha} - \alpha_0)}{\sigma_0} \xrightarrow{d} \mathrm{N}\left(0, 1\right), \quad \text{as } n \to \infty, \quad \sigma_0^2 := \frac{\mathrm{E}\left[\varepsilon^2 \nu^2\right]}{\left(\mathrm{E}\left[\nu^2\right]\right)^2}.$$

... *even with* p *(much) greater than* n!

$\Rightarrow$ Normal approximation valid even in high-dim. regime.

REF: Belloni, Chernozhukov, Hansen [2014 ReStud, EconPersp].

- Changed field of econometrics!

Numerical Illustration

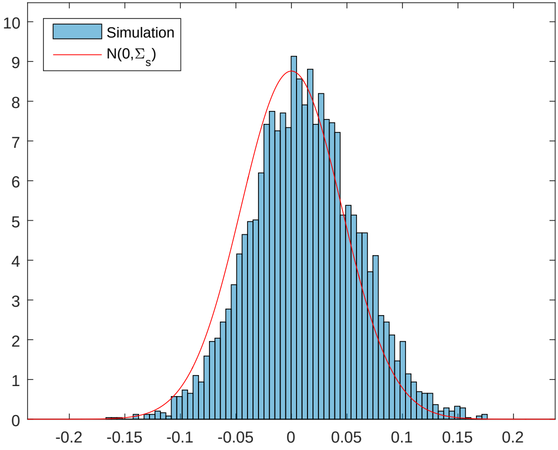


Figure: Post-Double Lasso $\sqrt{n}(\check{\alpha} - \alpha_0)$ vs. N $\left(0, \sigma_0^2\right)$

Variance Estimation

For $\sqrt{n}(\check{\alpha} - \alpha_0)/\sigma_0 \xrightarrow{d} \mathrm{N}(0,1)$ useful need to estimate

$$\sigma_0^2 = \frac{\mathrm{E}\left[\varepsilon^2 \nu^2\right]}{\left(\mathrm{E}\left[\nu^2\right]\right)^2}.$$

Analogy principle suggests:

$$\check{\sigma}^2 := \frac{n^{-1}\sum_i \widehat{\varepsilon}_i^2 \widehat{\nu}_i^2}{\left(n^{-1}\sum_i \widehat{\nu}_i^2\right)^2},$$

$$\text{where} \quad \widehat{\varepsilon}_i := Y_i - \widehat{\alpha} D_i - Z_i' \widehat{\gamma} \quad \text{and} \quad \widehat{\nu}_i := D_i - Z_i' \widehat{\psi}.$$

Under regularity conditions, Post-Double Lasso satisfies

$$\frac{\sqrt{n}(\check{\alpha} - \alpha_0)}{\check{\sigma}} \xrightarrow{d} \mathrm{N}(0,1).$$

Confidence Interval with Post-Double Lasso

$\xi \in (0,1)$: Significance level (e.g. $\xi = .05$)

$q_\xi := \Phi^{-1}(\xi)$: $\mathrm{N}(0,1)$ quantile function (e.g. $q_{.025} = 1.96$)

Then
$$\mathrm{P}\left(\alpha_0 \in \left[\check{\alpha} \pm q_{1-\xi/2}\frac{\check{\sigma}}{\sqrt{n}}\right]\right) \to 1 - \xi.$$

Define $100 \times (1-\xi)\,\%$ confidence interval (CI):

$$\check{\mathrm{C}}\mathrm{I}(1-\xi) := \left[\check{\alpha} \pm q_{1-\xi/2}\frac{\check{\sigma}}{\sqrt{n}}\right].$$

Asymptotically valid—even in high-dim. regime!

Post-Double Lasso as Feasible IV

Estimator

$$\check{\alpha} = \frac{\sum_{i=1}^{n}(D_i - Z_i'\widehat{\psi})(Y_i - Z_i'\widehat{\gamma})}{\sum_{i=1}^{n}(D_i - Z_i'\widehat{\psi})D_i}.$$

IF we knew γ_0 and ψ_0, we observe

$$\widetilde{Y}_i := Y_i - Z_i'\gamma_0 \qquad \text{('outcome')}$$

$$\widetilde{D}_i := D_i - Z_i'\psi_0 \qquad \text{('instrument')}$$

$\widetilde{D}$ function of $X = (D, Z')'$, so $E[\varepsilon \widetilde{D}] = 0$.

Suggests

$$\widetilde{\alpha}^{\mathtt{IV}} := \frac{\sum_i \widetilde{D}_i \widetilde{Y}_i}{\sum_i \widetilde{D}_i D_i}.$$

$\check{\alpha}$ operationalizes this idea.

Orthogonalized Moments

A Moment Approach

From $E[Y|D,Z] = \alpha_0 D + Z'\gamma_0$ we see $(\alpha_0, \gamma_0')'$ solves

$$E\left[(Y - \alpha_0 D - Z'\gamma_0)\begin{pmatrix} D \\ Z \end{pmatrix}\right] = \mathbf{0}.$$

Moment condition. Starting point of estimation.

α_0 of interest. γ_0 pure nuisance.

γ_0 long $\Rightarrow$ possibly very noisy (biased) estimate.

Want moment condition for α_0 which is ‘insensitive’ to error in γ_0.

Orthogonalized Moments, I

$$E\left[\left(Y-\alpha_0 D-Z'\gamma_0\right)\begin{pmatrix} D \\ Z \end{pmatrix}\right]=\mathbf{0},$$

Consider (other) moment condition for α_0:

$$E\left[\left(Y-\alpha_0 D-Z'\gamma_0\right)\left(D-Z'\psi_0\right)\right]=0.$$

Has following zero derivative property:

$$\frac{\partial}{\partial\gamma_0}E\left[\left(Y-\alpha_0 D-Z'\gamma_0\right)\left(D-Z'\psi_0\right)\right]=E\left[\left(-Z\right)\left(D-Z'\psi_0\right)\right]=\mathbf{0}.$$

Moment orthogonalized wrt. γ_0.

Interpret: (Limited) nuisance estimation error has little impact.

Orthogonalized Moments, II

But we introduced new (nuisance) parameters ψ_0.

So how did we progress?

Luckily, by choice of moment condition

$$\frac{\partial}{\partial \psi_0} E\left[\left(Y - \alpha_0 D - Z'\gamma_0\right)\left(D - Z'\psi_0\right)\right] = E\left[\cancel{(-Z)\left(D - Z'\psi_0\right)}\right] = \cancel{0}$$

$$E[(Y - \alpha_0 D - Z'\gamma_0)(-Z)] = E[\varepsilon(-Z)] = \underline{\underline{0}}$$

Another zero derivative. Also orthogonalized wrt. ψ_0.

Constructing/exploiting such zero derivatives active research topic.

Other Methods for High-Dimensional Regression

Other Methods

Our focus: Lasso.

- In part due to (solid) theoretical foundation.
- In part due to popularity.

Other high-dim. methods exist.

Could take the place of Lasso in (most of) the above.

- à la "Post-Double X"

Dantzig Selector

Dantzig Selector, I

To develop intuition, recall OLS:

$$\widehat{\beta} = \underset{b \in \mathbf{R}^p}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2.$$

Corresponding FOCs:

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' \widehat{\beta}) X_{ij} = 0 \quad \text{for all } j = 1, \ldots, p$$

Lasso changes criterion:

$$\widehat{\beta}(\lambda) = \underset{b \in \mathbf{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2 + \lambda \|b\|_1 \right\}.$$

Alternatively: Modify FOCs.

Dantzig Selector, II

Dantzig Selector (DS)

$$\widehat{\beta}(\lambda) = \operatorname*{argmin}_{b \in \mathbf{R}^p} \|b\|_1$$
$$\text{s.t.} \quad \left| \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b) X_{ij} \right| \leqslant \lambda \text{ for all } j = 1, \ldots, p$$

Thus, we relax

- Ensure OLS FOCs
- Encourage sparsity (minimize ℓ_1-norm)

DS important because of straightforward IV extension.

REF: Candes & Tao (2007), "The Dantzig selector: statistical estimation when p is much larger than n" *Annals of Statistics*

Ridge Regression

Ridge Regression

$$\widehat{\beta}(\lambda) = \underset{b \in \mathbf{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2 + \lambda \|b\|_2^2 \right\}$$

Akin to Lasso: Replaces ℓ_1 penalty $\|b\|_1$ with ℓ_2 penalty $\|b\|_2^2$

Explicit solution:

$$\widehat{\beta}(\lambda) = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i' + \lambda \mathbf{I}_p \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \right)$$

Ridge does <u>not</u> perform variable selection ($x \mapsto x^2$ flat around zero)

Lasso now more popular because of automatic variable selection.

Shrinkage: Orthonormal Design, I

With $n^{-1}\sum_i X_i X_i' = \mathbf{I}_p$, Ridge solution

$$\widehat{\beta}_j^{\texttt{Ridge}}(\lambda) = \frac{\widehat{\beta}_j^{\texttt{LS}}}{1+\lambda}, \quad j = 1, 2, \ldots, p.$$

Proportional shrinkage.

Recall soft-thresholding:

$$\widehat{\beta}_j^{\texttt{Lasso}}(\lambda) = \mathrm{sgn}(\widehat{\beta}_j^{\texttt{LS}})\left(|\widehat{\beta}_j^{\texttt{LS}}| - \frac{\lambda}{2}\right)_+, \quad j = 1, 2, \ldots, p.$$

Amounts to fixed shrinkage.

Shrinkage: Orthonormal Design, II

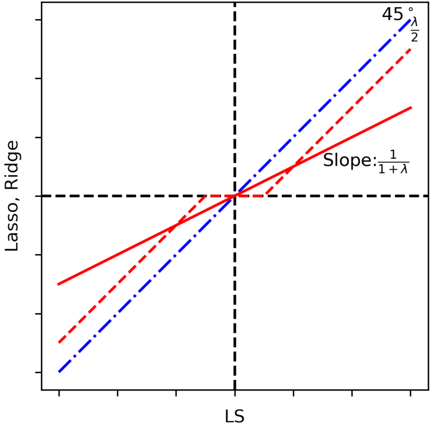


Figure: Ridge and Lasso vs. Least Squares

Implementing Ridge in Python

```python
import numpy as np
from sklearn import datasets
from sklearn.linear_model import Ridge
boston = datasets.load_boston()
X = boston.data
y = boston.target
fit = Ridge(alpha = 1).fit(X,y) # alpha = penalty
y_pred = fit.predict(X)
coef = fit.coef_
print(np.round(coef,2))
```

Cross-Validation and Ridge

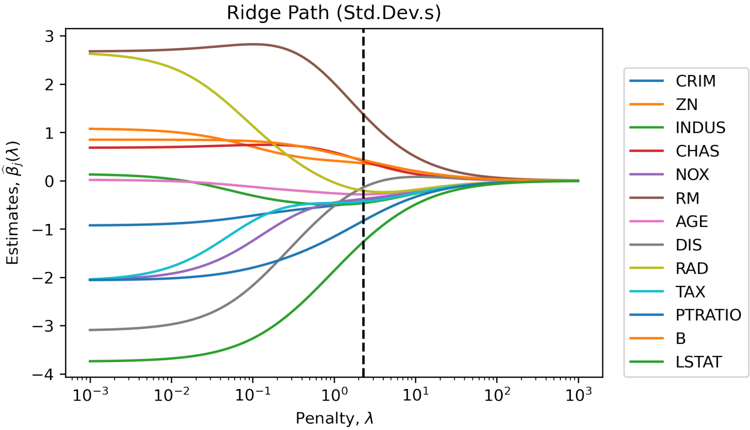
Ridge penalty typically determined by sample splitting/cross-validation

- Implementation and discussion analogous to Lasso

To implement Ridge with cross-validation in Python:

1. import `RidgeCV` instead
2. and replace `Ridge(alpha = 1)` with `RidgeCV(cv = 5)`

Ridge Path with Basic Boston Housing Data



Vertical line = CV penalty.

Elastic Net

Elastic Net

Elastic Net: Somewhere in between Lasso and Ridge:

$$\widehat{\beta}(\lambda, \ell) := \underset{b \in \mathbf{R}^p}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - X_i' b)^2 + \lambda \left[\ell \, \|b\|_1 + (1 - \ell) \, \|b\|_2^2 \right] \right\}.$$

Idea: When some regressors highly correlated, Lasso may perform poorly.

- "A bit of Ridge" provides stability.
- Orthonormal case: Part fixed/proportional shrinkage.

Elastic Net in Python

```python
# Basic implementation
from sklearn.linear_model import ElasticNet
fit=ElasticNet(alpha=1,l1_ratio=0.1).fit(X,y)
```

May choose penalty parameters λ and ℓ via splitting/CV:

```python
from sklearn.linear_model import ElasticNetCV
fit = ElasticNetCV(cv = 5).fit(X,y)
```

Normalization warning still applies.

Where are we going?

Part	Topic	Parameterization non-linear	Estimation non-linear	Dimension	Numerical optimization	**M-estimation** (Part III)	Outcome (y_i)	Panel (c_i)
I	OLS	÷	÷	low	÷	✓	$\mathbb{R}$	✓
II	**LASSO**	÷	✓	high	✓	÷	$\mathbb{R}$	÷
IV	Probit	✓	✓	low	✓	✓	$\{0, 1\}$	÷
	Tobit	✓	✓	low	✓	✓	$[0; \infty)$	÷
	Logit	✓	✓	low	✓	✓	$\{1, 2, ..., J\}$	÷
	Sample selection	✓	✓	low	✓	✓	$\mathbb{R}$ and $\{0, 1\}$	÷
	Simulated Likelihood	✓	✓	low	✓	✓	Any	✓
	Quantile Regression	÷	✓	(low)	✓	✓	$\mathbb{R}$	÷
	Non-parametric	✓	(✓)	∞	÷	÷	$\mathbb{R}$	÷